DOCUMENT RESUME

ED 260 560 EC 180 527

AUTHOR Scruggs, Thomas E.
TITLE The Administration and Interpretation of Standardized Achievement Tests with Learning Disabled and Behaviorally Disordered Elementary School Children. Year Two Final Report.
INSTITUTION Utah State Univ., Logan.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE 15 Jul 85
GRANT G008300008
NOTE 323p.; Prepared by the Developmental Center for Handicapped Persons. For earlier report, see ED 256 082.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC13 Plus Postage.
DESCRIPTORS *Achievement Tests; *Behavior Disorders; Elementary Education; *Learning Disabilities; Standardized Tests; Teaching Methods; *Test Wiseness

ABSTRACT

Several experiments were carried out to determine whether learning disabled (LD) and behaviorally disordered (BD) students exhibit deficiencies in appropriate test-taking strategies and, if so, whether these strategies could be successfully trained. Preliminary investigations indicated that mildly handicapped students do exhibit deficiencies in this area, including attention to inappropriate distractors, failure to successfully employ prior knowledge and deductive reasoning strategies, and failure to identify correctly specific types of questions which call for different strategies. Deficiencies were also observed regarding use of separate answer sheets and expressed attitudes toward tests. In year 1, approximately 100 LD and BD elementary (grades 2-4) were randomly assigned to treatment (training on test-taking skills) or control conditions. All Ss scored significantly higher on a test of test-taking skills. During year 2, approximately 100 LD and BD Ss (grades 4-6) were randomly assigned to treatment (training involving both reading and math subtest areas of the Stanford Achievement Test). Trained Ss scored significantly higher on two subtests and descriptively higher on a third subtest. Extensive appended material includes 19 items (journal articles, conference papers, and manuscripts unpublished or submitted for publication) on test-taking skills and their implications for LD and BD students. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

BEST COPY AVAILABLE

ED260560

THE ADMINISTRATION AND INTERPRETATION OF STANDARDIZED ACHIEVEMENT TESTS WITH LEARNING DISABLED AND BEHAVIORALLY DISORDERED ELEMENTARY SCHOOL CHILDREN

(Grant No. G008300008)

YEAR TWO FINAL REPORT

Submitted to

Special Education Programs
(CFDA 84.023C)
U.S. Department of Education,
Office of Special Education

July 15, 1985

Dr. Thomas E. Scruggs (801) 750-1224
Developmental Center For Handicapped Persons
Utah State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Thomas E.
Scruggs
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Table of Contents

Page

Abstract 1

PROJECT OVERVIEW 3

Year One Activities 3
1. Assessment of spontaneously employed test-taking strategies 3
2. Development and revision of training materials ... 9

Year Two Activities 14
1. Teacher validation for training materials 14
2. Needs assessment 14
3. Materials development 15
4. Pilot test 16
5. Field test 16
6. Experimental study 16
7. Other accomplishments 17

Publications 19

Presentations 22

Manuscripts 24

Unpublished Products 26

Appendix A:
Research in Progress: Improving the Test-Taking Skills of LD and BD Elementary Students

Appendix B:
An Analysis of Children's Strategy Use on Reading Achievement Tests

Appendix C:
Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children

Appendix D:
Passage Independence in Reading Achievement Tests: A Follow-Up

Appendix E:
Are Learning Disabled Students "Test-Wise?": An Inquiry into Reading Comprehension Test Items

Appendix F:
Learning Disabled Students' Spontaneous Use of Test-Taking Skills on Reading Achievement Tests

Table of Contents (continued)

Appendix G:
Attitudes of Behaviorally Disordered Students Toward Tests

Appendix H:
Format Changes in Reading Achievement Tests: Implications for Learning Disabled Students

Appendix I:
Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis

Appendix J:
Improving the Test-Taking Skills of Learning-Disabled Students

Appendix K:
The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children

Appendix L:
Improving the Test-Taking Skills of Behaviorally Disordered and Learning Disabled Children

Appendix M:
Can LD Students Effectively Use Separate Answer Sheets?

Appendix N:
Attitudes of Behaviorally Disordered Students Toward Tests: A Replication

Appendix O:
The Effects of Coaching on the Standardized Test Performance of Mildly Handicapped Students

Appendix P:
Current Conceptions of Test-Wiseness: Myths and Realities

Appendix Q:
Academic and Intellectual Characteristics of Behaviorally Disordered Children and Youth

Appendix R:
Academic Characteristics of Behaviorally Disordered and Learning Disabled Students

Table of Contents (continued)

Appendix S:
Improving the Test-Taking Skills of Learning Disabled Students

Appendix T:
SUPER SCORE: Training manuals and workbooks for the Stanford Achievement Test

Appendix U:
SUPER SCORE III: Training manuals and workbooks for the Iowa Test of Basic Skills

Appendix V:
SUPER SCORE II: Training manuals and workbooks for the Comprehensive Test of Basic Skills

Appendixes T, U, and V were not included in the copy received by ERIC; copyrighted materials.

Abstract

Several experiments were carried out over the course of a 24-month period to determine whether: (a) learning disabled (LD) and behaviorally disordered (BD) students exhibit deficiencies with respect to appropriate test-taking strategies, and, if so, (b) whether these strategies could be successfully trained. Preliminary investigations indicated that mildly handicapped students do exhibit deficiencies in the area of test-taking strategies. These deficiencies include attention to inappropriate distractors, failure to successfully employ prior knowledge and deductive reasoning strategies, and failure to identify correctly specific types of questions which call for different strategies. In addition, deficiencies were observed with respect to use of separate answer sheets and expressed attitudes toward tests. In the first year test-training evaluation, approximately 100 LD and BD elementary-age students representing grades 2, 3, and 4 were randomly assigned to treatment and control conditions. Treatment subjects received eight training sessions on test-taking skills with particular regard to the Stanford Achievement Test (SAT). All students scored significantly higher on a test of test-taking skills. In addition, third and fourth grade LD and BD students scored significantly higher on the Word Study Skills subtest and exhibited descriptive increases over the experimental group with respect to other subtests. Second grade students were apparently

unaffected by the training procedure. In addition, a similar test-training package applied to intact third grade classrooms of mostly nonhandicapped students indicated that these materials were successful in improving student attitudes toward the test-taking experience.

During the Year 2 test training evaluation, approximately 100 LD and BD fourth, fifth, and sixth grade students were randomly assigned to treatment and control conditions. Treatment condition subjects received five days of training on revised and extended training materials which involved both reading and math subtest areas of the SAT. Results indicated that trained students scored significantly higher on two subtests and descriptively higher on a third subtest. In a second experiment, 24 special education teachers (of approximately 200 students) were assigned at random to training and control conditions. Training condition teachers were given materials for five days of training of test-taking skills for the Iowa Test of Basic Skills (ITBS). Data from this investigation will be analyzed during Year 3 of the project.

PROJECT OVERVIEW

The primary objective of this project was to determine whether scores on standardized achievement tests could be improved through a combination of reinforcement, practice, and training of "test-taking skills"; that is, those skills which refer to understanding of the most efficient means to take a test rather than knowledge of the content area (see "Research in Progress," Appendix A). Such training, if successful, would likely improve the validity of resulting test scores in that a potential source of error, i.e., difficulty with format, testing conditions, etc., would be eliminated. In addition to the major objectives, several smaller investigations were planned and carried out, the ultimate objective of which was to determine whether, in fact, students in special education placement exhibited specific deficiencies on selected aspects of test-taking.

Year One Activities

A series of studies was initiated to evaluate what specific skills lower functioning students may lack with respect to test taking, and to develop a new set of materials which might address these needs. Accomplishments are described below by each task.

1. Assessment of spontaneously employed test-taking strategies (July-December, 1983). A shorter version of the Stanford Achievement Test, Reading subtests, questionnaire form and follow-along sheet, was developed in order to evaluate the

skills students spontaneously employed in test-taking situations. These materials were utilized in several studies to acquire this information. Students were selected from two remedial and one original program from each of grades 1 through 7. Students were individually administered selected subtests of the Stanford Achievement Test. They were asked for their level of confidence for each answer and the strategies they had chosen for answering the questions. It was determined that a complete hierarchy of strategies existed with respect to answering test questions beyond simply knowing or not knowing the answer, and that these strategies resulted in differential levels of performance on the part of the students. This investigation is described in detail in the manuscript in Appendix B entitled, "An Analysis of Children's Strategy Use on Reading Achievement Tests". This manuscript has been published in Elementary School Journal. Additional evaluation of the data from this investigation indicated the existence of a developmental trend through the elementary grades in the use of elimination strategies on ambiguous multiple choice items. That is, as children got older they became more proficient with respect to their spontaneous ability to eliminate inappropriate or obviously incorrect alternatives. These results have also been described in detail in the manuscript entitled, "Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children," which is given in Appendix C.

A test of "passage independence" of reading comprehension test items on the Stanford Achievement Test was developed by administering items from the Reading Comprehension subtest of the SAT to college undergraduates. The purpose of this investigation was to determine what proportion of these test items were potentially answerable by employing prior knowledge or deductive reasoning skills. It was determined that college undergraduates were able to answer nearly 80% of these questions on the average, with many students answering them all correctly. This article is given in Appendix D under the title, "Passage Independence in Reading Achievement Tests: A Follow-Up," and has been published in the journal Perceptual and Motor Skills.

Two follow-up investigations were intended to examine more precisely the nature of test-taking strategies employed by learning disabled students, specifically as compared with the strategies employed by their non-disabled counterparts. In one investigation, LD and non-LD students were administered items from the Stanford Achievement Test, Reading Comprehension subtest, with the actual reading passages deleted from the test. Students were told to simply answer the questions the best that they could. In the second experiment, all items were read to both groups of students in order to control for general reading ability. In both experiments, students not classified as learning disabled scored significantly higher on this test of "passage independent" test items than did their learning disabled counterparts. These results indicated (a) that learning disabled students may differ with respect to spontaneous test-taking strategies, such as use of prior knowledge and deductive reasoning skills, and (b) raise the issue of what such test items are actually measuring, since they could be so easily answered without having read the corresponding passage. This investigation has been written in manuscript form and is in Appendix E under the title, "Are Learning Disabled Students Test-Wise: An Inquiry into Reading Comprehension Test Items." It has been submitted for publication and was presented at the annual meeting of the American Educational Research Association, Chicago, April, 1985* (see footnote, page 23).

In a second investigation, learning disabled and non-learning disabled students were directly questioned with respect to strategies they employed on reading comprehension test items and letter sounds test items. In this investigation, it was found that learning disabled students did not differ from their non-disabled peers with respect to answering recall comprehension questions, with ability to read controlled. However, learning disabled students were less likely to employ appropriate strategies to answer inferential questions and reported inappropriately high levels of confidence in their responses. In addition, when they did report using appropriate strategies, they were much less likely to employ them successfully. This project is described in detail in the manuscript, "Learning Disabled Students' Spontaneous Use of Test-Taking Skills on Reading Achievement Tests" (Appendix F). This manuscript has been accepted for publication in Learning Disability Quarterly and was presented at the annual meeting of the American Educational Research Association in New Orleans in April, 1984.

In a separate investigation it was determined that a sample of elementary-age behaviorally disordered students scored significantly lower than their nonhandicapped counterparts with respect to reported attitudes towards tests and the test-taking situation. This manuscript was published in the journal Perceptual and Motor Skills and is given in Appendix G. These investigations, taken together, provided valuable information regarding the most optimal training package to be developed for use with mildly handicapped students.

An evaluation of all major achievement tests was also made in order to determine whether tests were similar or different with respect to format demands on the test taker. In this investigation, all levels of six major achievement tests were evaluated for number of format changes per minute throughout the reading achievement test subtest. It was determined that achievement tests varied widely with respect to format demands, with most format changes occurring in the primary grades. These results are documented in the manuscript, "Format Changes in

Reading Achievement Tests: Implications for Learning Disabled Students," which can be found in Appendix H and has been accepted for publication in Psychology in the Schools.

In order to evaluate appropriately all previous attempts to train test-taking skills in the elementary grades, a meta-analysis was completed of all available studies in this area. It was determined that although the general effect of training was positive, differences in favor of training groups did not seem to become substantial unless training was relatively extensive. In addition, this meta-analysis revealed that low SES children and primary grade children were more likely to benefit from extended training hours. This seems to underline the importance in the present project of implementing a package with a higher level of intensity. The detailed results of this meta-analysis are given in Appendix I under the title, "Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis." This manuscript has been accepted for publication in Elementary School Journal.

Finally, during the first part of the project, the scope of the proposed research was described and published by Exceptional Children in the fall of last year and is given in Appendix A under the title, "Research in Progress: Improving the Test-Taking Skills of Learning Disabled and Behaviorally Disordered Elementary Students." In addition, during the fall, preliminary findings were reported at the seventh annual conference of Severe Behavior Disorders of Children and Youth in Tempe, Arizona, in a presentation entitled, "Training Behaviorally Disordered Children to Take Tests."

It was the intention of all of the above investigations to evaluate both tests and test-taking strategies of mildly handicapped students in order to determine the most likely strategies for intervention and the form that intervention should take. In all, it was determined that mildly handicapped students do differ from their nonhandicapped peers with respect to use of appropriate strategies on standardized achievement tests. It was also determined that these strategy deficits included use of prior knowledge, use of deductive reasoning skills, attention to appropriate distractors, and selection of strategies appropriate to correctly answering different types of terms.

2. Development and revision of training materials (September-February, 1983-1984). Based upon results of the above investigation and careful evaluation of the Stanford Achievement Test, materials were developed which were intended to teach to second, third, and fourth grade children in special education placements skills appropriate to the successful taking of the Stanford Achievement Test. These materials included eight scripted lessons and a student workbook of exercises on subtests meant to be very similar to those used on the Stanford Achievement

Test. These materials were intended to teach both general test-taking strategies, such as efficient time usage, as well as specific lessons meant to increase understanding of the particular test demands of the individual reading subtest of the Stanford Achievement Test. These materials are included with the Year 1 Final Report and corresponding ERIC Document and are entitled "Super Score."

Following the preliminary development of materials, they were pilot-tested in November on two groups of second grade children with learning and behavioral disorders. On the basis of this pilot investigation, several revisions were made in the materials. Specifically, some of the lessons proved to be too long, and some instructions were judged to be ambiguous. In addition, a pre- and posttest measure which was developed for use with this population was also judged to be inadequate to effectively assess progress made on these materials.

On the basis of the initial pilot investigation, the materials were revised and expanded to include second to fourth grades and were then implemented in a larger field test involving 16 students in special education placements in second and third grades. Students were randomly assigned to treatment and control groups at each of the three grade levels, and the lessons were administered to the treatment groups. Students in the experimental group were seen to score higher than students in the control group on a shortened version of the Stanford Achievement Test, Word Study Skills subtest. This investigation was reported in a manuscript which was published in Perceptual and Motor Skills, Appendix J.

Some final revisions were made of the training materials on the basis of the second field test, and materials were finally prepared for spring implementation immediately prior to districtwide standardized test administration. While final revisions were being made, individual schools were contacted to be involved in a larger experimental study intended to validate these materials. For this study, approximately 110 students enrolled in special education classes in grades 2, 3, and 4 in two different large elementary schools were selected and randomly assigned to treatment and control conditions. Four persons, including the principal investigator, took part in the two-week training period which was administered at the end of March. This training was administered in eight 20- to 30-minute sessions given from Monday to Thursday for each of two weeks immediately prior to districtwide test administration. At the same time, materials were developed intended to increase test-taking skills on the Comprehensive Test of Basic Skills and were administered in the school districts adjacent to Utah State University. This training package was implemented in local third grade classes in order to determine (a) whether these procedures were appropriate for whole-class administration, (b) whether the materials developed for the Stanford Achievement Test could be easily adapted to other tests, and (c) whether such training could be seen to have an impact upon test scores, attitudes, and time on task during test administration.

The results of the training on the Comprehensive Test of Basic Skills in the local third grade classes indicated that students' attitudes had, in fact, qualitatively improved as a result of the test training. It was suggested that the test training had resulted in a more normal distribution of attitudes after the end of the three days of testing and implied that the training had made the test-taking experience itself less traumatic on the part of third grade regular classroom students (including 15% mildly handicapped students). Time on-task during directions and during the test-taking experience itself did not seem to be affected by the training package. In addition, the training was seen to significantly increase the scores of students in the lower half of the class on the Word Attack subtest of the reading test. Analysis of the top half, or the group as a whole, was not possible due to the presence of strong ceiling effects in both experimental and control groups. This investigation has been written in manuscript form and is given in Appendix K under the title, "The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children."

Results of the training package with second, third, and fourth grade special education students also indicated that the training was successful in improving scores on standardized achievement tests. Although only descriptive differences were seen in some subtests, the training package significantly improved the performance of the experimental students over control students in the Word Study Skills subtest. This improvement was judged to be approximately equivalent to a three- to four-month increase in equivalent grade level. The fact that improvement in the Word Study Skills subtest was observed was considered to be due to the fact that this particular subtest involved many smaller subtests, several format changes, and potentially confusing directions for which the training package was thought to have been particularly helpful. Descriptive differences were seen in other subtests of the SAT but, not being statistically significant, it is not possible to determine whether they were a result of the training or simply sampling error. Evaluation of scores of the second grade students indicated that they apparently had not benefited from the training package. However, the differentially small number of subjects in the second grade sample, attrition suffered during the training, and the fact that the two 2nd grade groups were in retrospect found to have differed with respect to the previous year's testing, obscure clear interpretation of this data. It may be, for example, that second grade LD and BD students have insufficient reading and other academic skills to enable them to benefit from this training package, or it could be that these students had in fact benefited but that due to sampling and attrition problems these benefits were not observed. This entire investigation has been described in detail and is given in Appendix L under the title, "Improving the Test-Taking Skills of Behaviorally Disordered and Learning Disabled Children," which has been accepted for publication in Exceptional Children.

Year Two Activities

Progress of the second year's activities has proceeded in accordance with the planned schedule of activities. These activities are described below:

1. Teacher validation for training materials (July through June 1984-1985). Materials developed during Year 1 were further adapted for teacher use for the Iowa Test of Basic Skills (see Appendix U) and given to a randomly assigned experimental group of special education teachers (N = 24) in Mesa, Arizona, for implementation during the two weeks immediately prior to yearly testing. This training took place in April, 1985. When test data become available, test scores will be compared statistically with the control group.

2. Needs assessment. Since a major format change in standardized tests, which takes place in the upper elementary grades, is the use of separate answer sheets, a preliminary

evaluation was made of the relative ability of learning disabled students to utilize separate answer sheets. Results of this investigation indicated that LD students differed with respect to speed of responding, but not accuracy of responding, with speed controlled. In addition, descriptive results suggested that LD students may be more likely to go outside the line of the answer circle. The manuscript which describes this investigation is entitled, "Can LD Students Effectively Use Separate Answer Sheets?" and is found in Appendix M. Additionally, a follow-up investigation to Year 1 was conducted on attitudes behaviorally disordered students report toward achievement tests. Although the findings of Year 1 were somewhat contradictory (see Appendix G), the Year 2 investigation provided additional information that BD students do express more negative attitudes toward testing. The manuscript describing this investigation is entitled, "Attitudes of Behaviorally Disordered Students Toward Tests: A Replication," and is in Appendix N. Findings from the above two investigations were considered in developing Year 2 training materials.

3. Materials development (September/October, 1984). Based upon the results of a needs assessment, materials were developed to teach specific test-taking skills on reading and mathematics achievement tests and are given in Appendix T. Information gained from the development of materials from Year 1 was utilized. Since the Year 1 studies did not result in improvement on reading

comprehension subtests, training in this area was intensified.

4. Pilot test (November/December, 1984). As the materials were developed, they were pilot-tested on a small group of children in order to determine whether they do, in fact, teach the skills which they are intended to teach.

5. Field test (November/December, 1984). This test was not conducted because of the early (February 1) test administration in the Granite District this year and the fact that pilot-testing results were satisfactory.

6. Experimental study (January/February, 1985). Based upon the results of the pilot test and the results of training from Year 1, an experimental study involving approximately 100 students in special education classes in grades four through six was implemented immediately prior to the regularly scheduled administration of district-wide tests, February 1. This training employed five 20- to 30-minute lessons with accompanying workbooks.

Test scores of experimental and control students were entered into a 2 (experimental vs. control) by 2 (LD vs. BD) analysis of variance on each of the five trained subtests. Results replicated those of Year 1 in that a significant effect was found for trained students on the Word Study Skills subtest. Trained students scored an average of 9 percentile points higher than untrained students, consistent with Year 1 findings, and considerably higher than many previous findings with non-handicapped students. In addition, a significant effect favoring trained students was found on the Mathematics Concepts subtest. An obtained interaction on this subtest indicated that training had exhibited a differential effect on behaviorally disordered students. In addition, a descriptive but non-significant effect favoring trained students was found on the Mathematics Computation subtest. As in Year 1, no effect was found for the Reading Comprehension subtest. This investigation is described in detail in the manuscript entitled, "The Effects of Coaching on the Standardized Test Performance of Mildly Handicapped Students," which is given in Appendix O.

7. Other accomplishments. A review paper critically evaluating "test-wiseness" and its implications for special education was written and accepted for publication in the journal School Psychology Review (Appendix P). Also, a meta-analysis of research on test-anxiety is being conducted. To date, 80% of available articles have been coded. A paper describing the utility of standardized achievement test scores was presented at the Conference on Severe Behavior Disorders, Tempe, Arizona, November, 1984* (see footnote, page 23). A manuscript based on this presentation has been accepted for publication in Behavioral Disorders Monographs and is in Appendix Q. Another paper describing differences between LD and BD students in achievement test scores, entitled "Academic Characteristics of Behaviorally

Disordered and Learning Disabled Students," has been tentatively accepted for publication in Behavioral Disorders and is in Appendix R. Finally, a presentation describing the project's activities was given at the annual meeting of the Association for Children and Adults with Learning Disabilities, San Francisco, February, 1985* (see footnote, page 23), and was attended by approximately 300 professionals. The paper from this project is in Appendix S.

Titles of project publications, presentations, manuscripts, and training materials generated to date are given on the following pages.

Publications

1. Lifson, S., Scruggs, T. E., & Bennion, K. (1984). Passage independence in reading achievement tests: A follow-up. Perceptual and Motor Skills, 58, 945-946. (Appendix D)
2. Mastropieri, M. A., Jenkins, V., & Scruggs, T. E. (in press). Academic and intellectual characteristics of behaviorally disordered children and youth. Monographs in Behavior Disorders, 9. (Appendix Q)
3. Scruggs, T. E. (in press). Administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school children. Final Report. Logan, UT: Utah State University. (ERIC Document Reproduction Service)
4. Scruggs, T. E., Bennion, K., & Lifson, S. (1985). An analysis of children's strategy use on reading achievement tests. Elementary School Journal, 85, 479-484. (Appendix B)
5. Scruggs, T. E., Bennion, K., & Lifson, S. (in press). Learning disabled students' spontaneous use of test-taking skills on reading achievement tests. Learning Disability Quarterly. (Appendix F)
6. Scruggs, T. E., & Lifson, S. A. (in press). Current conceptions of test-wiseness: Myths and realities. School Psychology Review, 14(3). (Appendix P)

7. Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled students. *Exceptional Children*. (Appendix L)

8. Scruggs, T. E., Mastropieri, M. A., Tolfa, D., & Jenkins, V. (1985). Attitudes of behaviorally disordered students toward tests. *Perceptual and Motor Skills*, *60*, 467-470. (Appendix G)

9. Scruggs, T. E., & Tolfa, D. (1985). Improving the test-taking skills of learning disabled students. *Perceptual and Motor Skills*, *60*, 847-850. (Appendix J)

10. Scruggs, T. E., White, K. R., & Bennion, K. (in press). Teaching test-taking skills to elementary grade students: A meta-analysis. *Elementary School Journal*. (Appendix I)

11. Scruggs, T. E., & Williams, N. J. (in press). *Teaching test-taking skills to learning disabled and behaviorally disordered children. SUPER SCORE: Test taking manual and workbooks*. Logan, UT: Utah State University. (ERIC Document Reproduction Service)

12. Taylor, C., & Scruggs, T. E. (1983). Research in progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary students. *Exceptional Children*, *50*, 277. (Appendix A)

13. Tolfa, D., Scruggs, T. E., & Bennion, K. (in press). Format changes in reading achievement tests: Implications for learning disabled students. Psychology in the Schools. (Appendix H)

Presentations

1. *Scruggs, T. E. (1985, February). Improving the test-taking skills of learning disabled students. Paper presented at the annual meeting of the Association for Children and Adults with Learning Disabilities, San Francisco, CA. (Appendix S)
2. Scruggs, T. E., Bennion, K., & Lifson, S. (1984, April). Spontaneously employed test-taking strategies of high and low comprehending elementary school children. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (Appendix B and F)
3. *Scruggs, T. E., & Lifson, S. A. (1985, April). Are learning disabled students 'testwise'? An inquiry into reading comprehension test items. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (Appendix E)
4. *Scruggs, T. E., & Mastropieri, M. A. (1984, November). Academic characteristics of behaviorally disordered and learning disabled students. Paper presented at the eighth annual conference on Severe Behavior Disorders of Children and Youth, Tempe, AZ.

5. Scruggs, T. E., & Taylor, C. (1983, November). Training behaviorally disordered children to take tests. Paper presented at the seventh annual conference on Severe Behavior Disorders of Children and Youth, Tempe, AZ.

---

*Out-of-state travel funds were not awarded this project for the 1984-85 funding year. However, since it is the view of the principal investigator that national presentations are of critical importance for immediate and widespread dissemination of project findings, alternate sources of funding were located to meet these expenses.

Manuscripts

1. Scruggs, T. E., Bennion, K., & Williams, N. J. (1985). *The effects of training in test-taking skills on test performance, attitudes, and on-task behavior of elementary school children*. Unpublished manuscript, Utah State University, Logan, UT. (Appendix K)
2. Scruggs, T. E., & Lifson, S. A. (1984). *Are learning disabled students 'test-wise?': An inquiry into reading comprehension test items*. Manuscript submitted for publication. (Appendix E)
3. Scruggs, T. E., & Mastropieri, M. A. (1984). *Academic characteristics of behaviorally disordered and learning disabled students*. Manuscript submitted for publication (accepted pending revisions, *Behavioral Disorders*). (Appendix R)
4. Scruggs, T. E., Mastropieri, M. A., & Tolfa, D. (1985). *The effects of coaching on the standardized test performance of mildly handicapped students*. Unpublished manuscript, Utah State University, Logan, UT. (Appendix O)
5. Scruggs, T. E., & Tolfa, D. (1985). *Developmental aspects of test-wiseness for absurd options: Elementary school children*. Unpublished manuscript, Utah State University, Logan, UT. (Appendix C)

6. Tolfa, D., & Scruggs, T. E. (1985). *Can LD students effectively use separate answeer sheets?* Manuscript submitted for publication. (Appendix M)

7. Tolfa, D., Scruggs, T. E., & Mastropieri, M. A. (1985). *Attitudes of behaviorally disordered students toward tests: A replication*. Manuscript submitted for publication. (Appendix N)

Unpublished Products

1. Scruggs, T. E. (1985). SUPER SCORE II: Training manuals and workbooks for the Comprehensive Test of Basic Skills. Logan, UT: Utah State University. (Appendix V)
2. Scruggs, T. E. (1985). SUPER SCORE III: Training manuals and workbooks for the Iowa Test of Basic Skills. Logan, UT: Utah State University. (Appendix U)
3. Scruggs, T. E. (1985). SUPER SCORE: Training manuals and workbooks for the Stanford Achievement Test. (Appendix T)

APPENDIX A

RESEARCH IN PROGRESS

Charles C. Cleland
Department Editor

Improving the Test-Taking Skills of LD and BD Elementary Students

Principal Investigators: Cle Taylor and Thomas Scruggs, Exceptional Child Center, Utah State University.

Purpose/Objectives: The purpose of this investigation is to determine whether reinforcement techniques and direct training in test-taking skills can increase the validity of test scores for learning disabled (LD) and behaviorally disordered (BD) students. To determine the degree to which LD and BD students exhibit inappropriate (inefficient) test-taking skills, students are observed and interviewed while taking standardized tests. Based on those observational data, procedures and training packages will be designed to increase student performance on standardized achievement tests. If the procedures and training are effective, educational decisions, which are frequently based in part on the results of standardized achievement tests, will be more valid because problems in areas such as test-taking skills, student motivation, and confusion due to testing format will be reduced or eliminated.

Subjects: Subjects are 100 elementary students enrolled in 12 resource rooms and self-contained classrooms for children with learning disabilities and behavioral disorders.

Methods: LD and BD children matched on age, handicap, and standardized achievement test score will be randomly assigned to experimental and control groups. Students in the experimental group will receive materials and procedures designed to improve the ability of handicapped students to take tests. Experimental and control groups will be compared statistically on several measures, including attitudes toward test-taking, student and teacher behavior during test administration, and actual performance on standardized tests of reading achievement. In following years, materials will be developed and implemented for mathematics achievement tests and test-taking skills for secondary-age handicapped students.

Results to Date: Preliminary findings indicate that many LD and BD children, as well as low achieving nonhandicapped students, do not spontaneously exhibit efficient test-taking behaviors. Specifically, handicapped children have been seen to exhibit difficulties with item format and distractors more typical of naive test takers.

Commencement and Estimated Completion Dates: This investigation began July 1, 1983 and is expected to continue for three years.

Funding: Funding for this investigation has been provided by a grant from the U.S. Department of Education, Research in Education of the Handicapped.

Publications/Products Available: Preliminary materials for improving test-taking skills, piloted on nonhandicapped second-grade students, have been developed and will be revised for use with handicapped children during the coming year. Manuscripts documenting the investigation will be completed and submitted for publication during the second half of the academic year. Please write the authors for further information.

"Research in Progress" is a forum for reporting ongoing research in the field of special education that has not yet been published. Investigators wishing to report studies in progress are invited to submit a brief synopsis of their efforts to the column editor, Charles C. Cleland, 3427 Monte Vista, Austin TX 78731. Reports are to be submitted in triplicate and should follow the format shown above, with a maximum length of 500 words.

APPENDIX B

An Analysis of Children's Strategy Use on Reading Achievement Tests

Thomas E. Scruggs
Karla Bennion
Steve Lifson
Utah State University

The Elementary School Journal
Volume 85, Number 4
© 1985 by The University of Chicago. All rights reserved.
0013-5984/85/8504-0002$01.00

Much of what constitutes reading instruction in today's public schools reflects students' scores on standardized achievement tests. Test performance may influence later assignment to reading groups, classrooms, or remedial or special education programs. Although norm-referenced reading tests have been criticized as insensitive to specific skill deficits and inadequate as complete diagnostic measures (Howell 1979), most reading tests have nonetheless been shown to be highly reliable and valid (Spache 1976). For better or worse, standardized reading tests are truly a part of education today and will most likely be used in the future.

If important decisions are to be based on the results of standardized reading tests, student scores should provide the best possible estimate of reading performance. Unfortunately, the results of past research indicate that reading test performance can be influenced by factors other than knowledge of test content (e.g., Taylor & White 1982). One of these factors, "test-wiseness" (TW), was first described in detail in 1965 by Millman, Bishop, and Ebel (p. 707) as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score." Millman et al. developed an outline of test-wiseness principles, which included time-using strategies, error-avoidance strategies, guessing strategies, and deductive-reasoning strategies. Slakter, Koehler, and Hampton (1970) presented information suggesting that TW has a developmental component. That is, students may become more "test-wise" as they grow older. Generally, researchers have inferred extent of TW on the basis of tests constructed specifically for this purpose.

Students themselves were questioned recently about strategies they use to answer test questions. Haney and Scott (1980) administered a number of achievement tests to 11 students, then questioned them the following day concerning how they attempted to answer each item. These researchers developed a complex model in which responses to interviewer questions were classified into 46 separate categories. Most of these categories included the use of some specific strategies such as guessing, elimination of alternatives, or "reasoning." Their results indicated that children use a wide range of strategies in answering test questions and that often a child's perception of item content bears little resemblance to the intentions of the test's author. Haney and Scott concluded that considerable "ambiguity" exists in standardized test questions, existing to a greater extent in science and social studies areas and to a lesser extent in reading areas.

Haney and Scott's work contributed significantly to our knowledge of the nature of ambiguous test items. However, the focus of their study was on test construction, with implications for the reduction of test item ambiguity. Although classroom teachers may use the results of Haney and Scott to improve their own tests, published standardized tests cannot be altered by teachers. A remaining question concerns the extent to which students employ test-taking strategies when faced with difficult or ambiguous items. Do students use such strategies spontaneously (that is, without being trained)? If so, which strategies (if any) are effective in obtaining correct answers? No previous research can be located to answer these questions.

To address these questions in the present study, the reading test performance of elementary school children was examined. Specifically, two areas were investigated: the strategies students spontaneously employed to answer reading test items and the relative effectiveness of these strategies in increasing reading test scores.

Procedure

A sample multiple-choice reading test based on items from the Stanford Achievement Test (SAT) (Madden, Gardner, Rudman, Karlsen, & Merwin 1973) was developed and piloted on five students to evaluate whether the length was appropriate and to establish reliable scoring conventions. This sample test included items from the Word Reading, Reading Comprehension, Word Study Skills, and Vocabulary subtests. After revisions had been made, it was administered to 31 elementary-age Caucasian students (15 girls, 16 boys) attending summer classes in a rural western area. Students were selected from both remedial and "enrichment" classes so a range of abilities was represented. As assessed by the Woodcock Reading Achievement Test (Woodcock 1973), 20 students read at or above grade level; 11 read below their grade level. Most students (20) were second or third graders, but students were also selected from Grades 1 (two students), 4 (two), 5 (five), and 6 (two).

All students were seen individually by one of four examiners. One examiner interviewed 18 students, whereas the other three interviewed two, four, and six students. First, students were given the Passage Comprehension subtest from the Woodcock Reading Achievement Test in order to identify an approximate reading comprehension grade equivalent. Students were then given selections from the SAT one year level higher than their assessed grade level on the Woodcock subtest. In this manner, a similar difficulty level was provided for each student. Most students were able to answer correctly approximately two-thirds of the test questions.

Students were then told to read aloud each test question (as well as the reading passages in the Reading Comprehension subtest) and whichever of the distractors they chose to read. They were neither encouraged nor discouraged from reading each distractor. As soon as students had answered a test question, they were asked to rate their level of confidence in their response: were they very sure, somewhat sure, or not sure the answer they had given was correct? After students had finished each subtest, they were asked to reread the questions and tell the examiner why they had chosen their answer. The examiner recorded reading errors, confidence levels, attention to distractors, reference to reading passages, and reported strategies. Sessions were tape-recorded to clarify any later ambiguity in scoring. Students spent 45–90 minutes in the session and answered 31–42 test questions. Some students received more questions than others because different levels of the SAT required different subtests and formats.

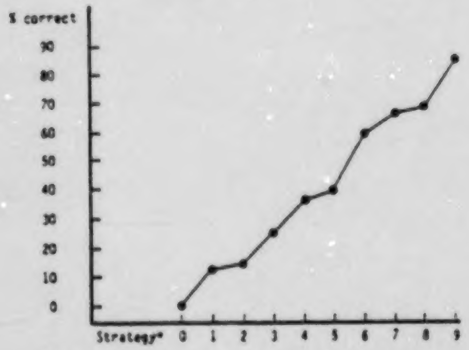


Fig. 1.—Percent correct answers by strategy used. Strategy classifications: 0, skipped item; 1, misread keyword; 2, faulty reasoning; 3, did not follow directions; 4, "seemed right;" 5, guessed; 6, used external evidence; 7, eliminated; 8, referred to passage; and 9, clearly "knew."

Results and discussion

Effectiveness of strategies

We found all strategy responses could be classified within a 10-level hierarchy that strongly predicted the probability of responding correctly. Proportions of correct responses were computed across subjects for each type of strategy and are shown in figure 1. These classifications were as follows: (*a*) skipped (student skipped the item), (*b*) misread a key word in question or distractors, (*c*) used faulty reasoning (example: one student reported, "This word must be the correct answer because it has a period after it"), (*d*) did not follow directions, (*e*) guessed, (*f*) "seemed right" (student thought the answer was correct without being able to state an explicit reason), (*g*) used external information (example: "I know most people in fires die from breathing smoke because a fireman told me that"), (*h*) eliminated inappropriate alternatives, (*i*) referred to passage, and (*j*) clearly "knew" the answer (example: "I know that a pear is a kind of fruit"). The existence of these strategies indicates that a complete hierarchy of test-taking skills exists beyond simply knowing or not knowing the answers, and these strategies can be more or less effective on a standardized reading test. For example, as seen in figure 1, when students skipped an answer, nothing was correct; when they guessed, they got 37% correct; when they eliminated alternatives, they got 67% correct. Proportions of employed strategies are given in table 1.

We condensed these strategies into five logical categories (skipping, procedural error, guessing strategy, deliberate strategy, and "knowing") and computed point-biserial correlations for each subject. The median correlation between item score and reported strategy was .54 ($p < .01$), a correlation of moderate strength.[1] No differential effects were seen by age, ability level, or examiner; although the sample was too small to conclusively investigate these possibilities.

Inspecting figure 1 reveals some other interesting findings. The high proportion of correct scores for guessing is notable. Since the number of answer choices varied between subtests and levels, with four choices the most common format, the probability of responding correctly by chance alone was estimated at .28. In fact, when students reported guessing, they

TABLE 1. Frequencies (*F*) and Percent (%) of Strategies Employed

Strategy level	*F*	%
0. Skipped item	9	1.0
1. Misread keyword	23	2.6
2. Faulty reasoning	38	4.3
3. Did not follow directions	7	.8
4. "Seemed right"	92	10.5
5. General	127	14.4
6. Used external evidence	21	2.4
7. Eliminated	45	5.1
8. Referred to passage	59	6.7
9. Clearly "knew"	458	52.1

scored 37% correct. "Guessing" responses scored virtually the same as "seemed right" responses, suggesting that even when students believe they are guessing, they still have some idea of what the correct answer might be and can use this strategy to advantage. "Seemed right" responses were common on the vocabulary subtests in which students often reported that a particular definition sounded correct but were otherwise uncertain. Another interesting finding is the high proportion of correct responses when the students reported using outside information or experience. Although content area tests, such as science and social studies, directly test outside knowledge, reading tests ostensibly are intended to test nothing besides knowledge of the passage's content. Therefore, although use of outside information should not help, students did benefit from the use of such information (however, when students referred to the passage, they scored even higher). The students' ability to use outside information as effectively as they did is surprising. This finding underlines the "passage independence" problems of reading comprehension items, a topic well investigated by researchers such as Tuinman (1973–74).

Level of confidence

Students had a reasonably good idea of whether they had answered a test question correctly. When students reported being "very sure" their answer was correct, they were correct 81% of the time. When they reported being "somewhat sure," they were correct only 13% of the time, and when they reported being "not sure," they obtained correct answers only 7% of the time. However, these figures are somewhat misleading. The results seem different if looked at another way: when students answered incorrectly, they also reported being "very sure" the answer was correct in 56% of the cases. Clearly, although related to performance, level of confidence in itself is not a sufficient check on correctness of a student's work. The relation between confidence and correctness of response was seen to vary widely from student to student, with a median point-biserial correlation of .29 ($p > .05$). Therefore, in many cases, other means are necessary for students to assess the correctness of their responses. These means will be described below.

The cost of carelessness

In addition to reported test-taking strategies, information was also collected on the degree to which the students attended to distractors and chose their answers by referring to the reading passage on the Reading Comprehension subtest. Results showed that students rarely referred to the reading passage; even though when they did, they stood a very good chance of answering the question correctly. In 89% of the cases where students answered a reading comprehension question incorrectly, they had not referred to the passage that clearly contained the correct answer. Of course, this does not mean that all of these questions could have been answered correctly had students referred to the passages, but it does appear that reading scores could be greatly improved by students' increased attention to the passages.

Similarly, a great deal of carelessness was observed in attention to distractors. When students answered incorrectly, in 40% of the 302 cases they had not read all distractors. Again, this finding does not mean all these questions could have been answered correctly by greater attention to distractors, but students could almost certainly have improved their scores by doing so. When students answered questions correctly, they had attended to all distractors in 73% of the 577 cases. It does appear, then, that test performance can be improved through greater attention to distractors.

Another surprising finding was the relatively small effect of reading errors. Although performance was clearly impaired when students misread a word of key importance (see fig. 1), in general misreading words was less detrimental than might be expected. When students misread one or more words in stem or distractor, the proportion of items answered correctly (58% of 293) was still quite high. Clearly, many students have developed strategies for coping with words they cannot read. It seems important to remind students not to "give up" if they cannot read every word. As the present investigation indicates, students are often able to answer correctly even though they cannot read every word.

One final finding concerning carelessness can be reported. All examiners noted the extent to which students had acted on the wrong stimulus in the "word study skills" subtest. In this subtest, students are given a word with an underlined sound and asked to find the same sound in one of three distractors. The following problem provides an example:

Prize
(*a*) prince
(*b*) size
(*c*) seven

The correct answer is *b* because the *z* in "size" has the same sound as the underlined *z* in "prize." What was surprising to us is that students often attended to the wrong stimulus, for example, the initial *pr* in the above question. Although the exact incidence of these errors cannot be given, their consistent occurrence seems to imply that teachers should stress the importance of attending to the underlined sound only.

Conclusions

The results of this study demonstrate that students do employ specific strategies to cope with test item ambiguity, indecision, or lack of knowledge in selecting correct answers. These findings have important implications directly bearing on student performance during testing. To attain the most correct answers, students should employ the strategies listed below:

1. Be certain to attend to all distractors and refer to the reading passage, even if you are "very sure" your answer is correct.
2. If you are having great difficulty reading a passage, read the questions and try to answer them anyway. Often, your own knowledge can help you choose an answer. If you have difficulty with some words in the question or distractors, answer anyway and base your answers on the words you can read.
3. If you have attended to all parts of a passage and test question and still do not know an answer, there is still a good chance of getting the correct answer if you guess.
4. Be certain you are attending to the appropriate stimulus, such as the underlined sound in a "word study skills" subtest. As in other subtests, wrong answer choices may look correct at first glance.

5. Make sure you answer every item. Even if you must hurry and guess frequently near the end, you will probably get some of the answers correct.

Considering the results of past research (Bangert, Kulik, & Kulik 1983), it is likely that to affect test performance significantly, a teacher will have to do more than simply read the above points to students. Examples and practice activities will help students develop these test-taking skills.

These findings should be of interest to special education teachers, particularly those in the area of learning disabilities. Many children are referred for special class placement on the basis of deficiencies in standardized reading-test scores. Special education often is quite beneficial to students who clearly need it, but before taking such a dramatic step, teachers should be certain that the test score reflects the best abilities of the student rather than a problem with test taking in general.

The present investigation indicates that a range of abilities exists in test-taking skills, as it does in other areas. If tests are to be as valid as possible, the specific skills observed in efficient students taking a reading test should be practiced by all students. If test-taking skills are incorporated in general test-administration procedures, it appears maximum benefit can be derived from the use of standardized reading tests.

Notes

The authors would like to thank Dr. Ginger Rhode and Judy Johnson, as well as Dr. Jay Monson, acting director, and the staff of the Edith Bowen School, particularly Dorothy Dobson and Lou Anderson, for their valuable assistance with this project. The authors would also like to thank Ursula Pimentel and Marilyn Tinnakul for typing the manuscript. Address requests for reprints to Thomas E. Scruggs, Exceptional Child Center, UMC 68, Utah State University, Logan, Utah, 84322.

[1] A point-biserial, rather than a Spearman correlation of ranks coefficient, was computed out of concern for the necessarily high number of ties resulting in computing a rank correlation with binary data. However, the obtained Spearman coefficient of .55 differed by only one point from the obtained point-biserial coefficient of .54.

References

Bangert, R. L.; Kulik, J. A.; & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. *Review of Educational Research*, **53**, 571–585.

Haney, W., & Scott, L. (1980). *Talking with children about tests: A pilot study of test item ambiguity* (National Consortium on Testing Staff Circular No. 7). Cambridge, MA: Huron Institute.

Howell, K. W. (1979). *Evaluating exceptional children*. Columbus, OH: Merrill.

Madden, R.; Gardner, E.; Rudman, H.; Karlsen, B.; & Merwin, J. (1973). *Stanford Achievement Tests*. New York: Harcourt, Brace, Jovanovich.

Millman, J.; Bishop, C. H.; & Ebel, R. (1965). An analysis of test wiseness. *Educational and Psychological Measurement*, **25**, 707–726.

Slakter, M. J.; Koehler, R. A.; & Hampton, S. H. (1970). Grade level, sex, and selected aspects of test-wiseness. *Journal of Educational Measurement*, **7**, 119–122.

Spache, G. (1976). *Identifying and diagnosing reading difficulties*. Boston: Allyn & Bacon.

Taylor, C., & White, K. R. (1982). The effects of reinforcement and training on group standardized test behavior. *Journal of Educational Measurement*, **19**, 199–210.

Tuinman, J. J. (1973–74). Determining the passage dependency of comprehension questions in five major tests. *Reading Research Quarterly*, **2**, 206–223.

Woodcock, R. W. (1973). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

APPENDIX C

Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children

Thomas E. Scruggs

Exceptional Child Center

Utah State University

Running head: Developmental Aspects

Abstract

Twenty-eight students from grades 1 through 5 were administered a test of test-wiseness for absurd options. Results suggested that a developmental trend may exist in test-wiseness for elementary-age school children.

Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children

First discussed by Thorndike in 1951, test-wiseness (TW) was described in detail by Millman, Bishop, and Ebel (1965), and defined as "a subject's capacity to utilize the characteristics and formats of the test and/or test-taking situation to receive a high score" (p. 707). They further described TW as "logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures" (Millman et al., 1965, p. 707). Ebel (1965) has suggested that error in measurement is more likely to be obtained from students low in test-taking skills. The student low in TW, therefore, may be more of a measurement problem than the student high in TW (Slakter, Koehler, & Hampton, 1970b).

Some investigations have indicated that TW has a developmental component; that is, that TW increases with age. Slakter, Koehler, and Hampton (1970a) administered a measure of TW to students from grades 5-11 and found a significant overall linear trend for grade level. Crehan, Koehler, and Slakter (1974) administered a TW test to students in grades 7 through 11, and a follow-up test to the same students two years later. Increases over all intervals except grades 9 to 11 were found. In a second follow-up of the same students, Crehan, Gross, Koehler, and Slakter (1978) replicated the previous findings and concluded that although TW increases by grade, large individual differences exist within grade levels.

Although the above investigations provide strong support for a developmental component of TW in the secondary grades, as yet no investigation has evaluated the developmental nature of TW in the elementary grades. The present investigation is intended to address this question.

Method

Subjects were 28 elementary school-age children attending summer classes prior to entering grades 1 through 5 in a western rural community. Students (1 first grader, 9 second graders, 11 third graders, 2 fourth graders, and 4 fifth graders) were selected from both remedial and "enrichment" classes so that a variety of ability levels was sampled.

Students were seen individually by one of four examiners. First, they were administered a five-item test of TW. This test was developed to measure the ability of students to eliminate options known to be incorrect (corresponding to the Millman et al., 1965 TW category I-D-1, absurd options). For example, one of the items was the following:

> Good airplane pilots must be able to _______
> quickly in an emergency.
>
> 1. fall asleep　　3. sturnate
> 2. scream　　4. thing

Students were orally provided with words they were unable to read. Since it was thought that evidence of TW would be more subtle in an elementary school population than it was in studies of secondary students, some departures were made from the procedures of Crehan et al. (1974). First, students were directly questioned regarding the reasons for their answer choices following completion of the test. Second, students were scored as reporting no elimination strategies (0), or reporting one or more strategies (1), regardless of the "correctness" of their answer to each test question.

Results and Discussion

A point-biserial correlation was computed between entering grade level of student and presence or absence of reported elimination strategies. The resulting coefficient, .44, was statistically significant ($p < .02$) and represented a moderate relation between grade level of student and reported use of elimination strategies, accounting for approximately 20% of total variance. Proportion of students reporting use of elimination strategies by grade level is given in Figure 1.

Insert Figure 1 about here

Thus, it appears that a developmental trend in one aspect of TW can be observed in children of elementary school age, and that this trend is similar to that seen in older students. These findings must be interpreted with caution, however, due to the limited sample size, as well as the fact that only one aspect of TW was measured. Although further research is needed, the results of this preliminary investigation suggest that students begin to learn TW skills as early as the primary grades, and that these skills continue to improve with age.

References

Crehan, K. D., Gross, L. J., Koehler, R. A., & Slakter, M. J. Developmental aspects of test-wiseness. Educational Research Quarterly, 1978, 3, 40-44.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. Longitudinal studies of test-wiseness. Journal of Educational Measurement, 1974, 11, 209-212.

Ebel, R. L. Measuring educational achievement. New Jersey: Prentice-Hall, 1965.

Millman, J., Bishop, H., & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 1970, 7, 119-122. (a)

Slakter, M. J., Koehler, R. A., & Hampton, S. H. Learning test-wiseness by programmed texts. Journal of Educational Measurement, 1970, 7, 247-254. (b)

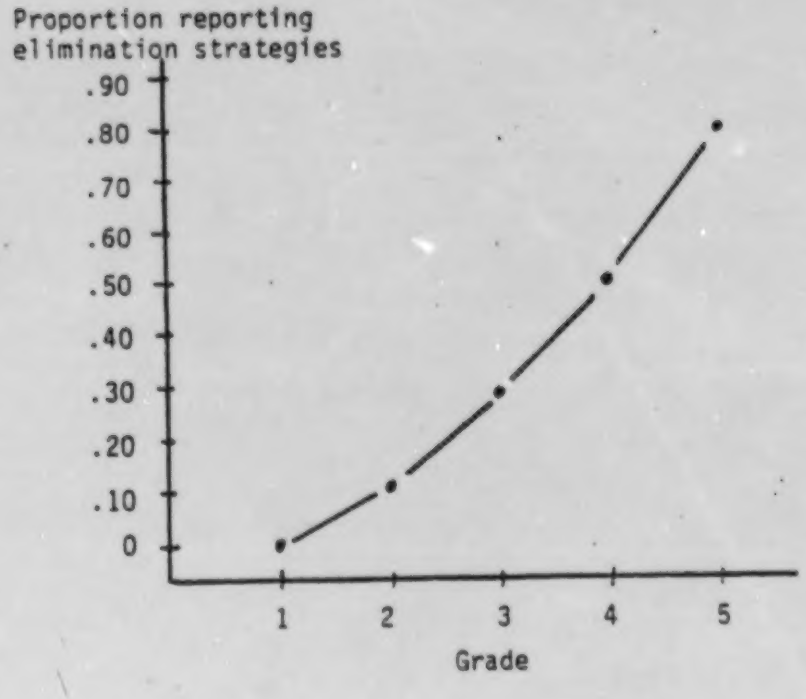
Thorndike, R. L. Reliability. In E. F. Linguist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951.

Footnote

[1] The author would like to thank Karla Bennion, Steve Lifson, Dr. Jay Monson and the staff of the Edith Bowen School for their assistance on this project.

Figure Caption

Figure 1. Proportion of students reporting elimination strategies by grade level.


Proportion reporting elimination strategies
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
1
2
3
4
5
Grade

APPENDIX D

Perceptual and Motor Skills, 1984, 58, 945-946. © Perceptual and Motor Skills 1984

PASSAGE INDEPENDENCE IN READING ACHIEVEMENT TESTS: A FOLLOW-UP[1]

STEVE LIFSON, THOMAS E. SCRUGGS, AND KARLA BENNION

Utah State University

Summary.—38 college undergraduates were administered reading-comprehension items from a major standardized achievement test with corresponding passages deleted. Analysis indicated that, after 20 years of similar research findings, highly passage-independent items still occur on major tests.

For almost 20 years, it has been documented that reading-comprehension test items can be answered correctly at above-chance rates without actually reading the relevant passage (Preston, 1964). Pyrczak (1976) mentions several types of items which seem particularly independent of the passage. These types include (a) items that can be answered from the examinee's own knowledge and (b) items about a particular passage that are related to each other in such a way that some items provide clues for other items. Reading-comprehension tests which include such items invite critical attention on the grounds that (a) examinees may have an advantage over those not using these strategies (Pyrczak, 1972) and (b), if a subject uses these principles and skips passages, he invalidates the purpose of the test (Tuinman, 1973-1974). Since an extensive review of the literature has shown no justification for the use of passage-independent items, the question arises as to whether these items still occur in commonly used standardized achievement tests. The present investigation was intended to determine whether such items are still in use.

METHOD

Subjects and Materials

Thirty-eight undergraduate elementary education students at a western university completed 16 multiple-choice reading-comprehension questions without the accompanying passages. The items selected were thought to represent questions that could be answered without having read the accompanying passage. These items were chosen to correspond to Millman, Bishop and Ebel's (1965) categories of test-wiseness strategies involving the general knowledge of the test taker and use of subject matter of neighboring items. The specific effects of these cues, however, were not addressed in this study. The 16 items were taken from the Stanford Achievement Test Form E, Level P-3, from a pool of 60 items. The items were kept in clusters illustrating which belonged together in terms of association with a particular passage.

---

[1] The authors thank Dr. Barnard Hayes for his kind and generous assistance with this investigation. Requests for reprints should be addressed to Steve Lifson, Exceptional Child Center, UMC 68, Utah State University, Logan, Utah.

Procedure

The materials were distributed to two sections of a class in teaching reading. The students were told: "Today I'm going to give you some reading-comprehension test items *without* the passages. It is not expected that you will answer all of the questions correctly; just do your best. Guess if you do not know the answer." No time limit was imposed upon the task.

RESULTS AND DISCUSSION

Analysis indicated that the mean score was 75% correct, with an average mean score of 11.9 of the 16 items. A one-sample t test (Hays, 1973) confirmed that the obtained scores were significantly different from chance responding ($t = 18.9, p < .001$).

Although the items were not randomly selected for this measure, they nevertheless represented 25% of the items included in the reading-comprehension section of the test. Clearly, at least some test developers have done little to alter passage-independent items in light of the research findings of almost two decades. While the effects of the readers' previous knowledge cannot be eliminated, the effects could be minimized by the use of fictional material for the passages with accompanying questions about the activities of an imaginary person. In spite of the reported validity of these items (SRA, 1979), the burden of construct validity rests with the authors of the tests. If some students are able to answer "reading-comprehension" test items correctly without reading the passage, one can question what is being measured.

REFERENCES

GARDNER, E. F., RUDMAN, H. C., KARLSEN, B., & MERWIN, J. C. *Stanford Achievement Test, Form E.* New York: Harcourt, Brace, Jovanovich, 1982.

HAYS, W. L. *Statistics for the social sciences.* New York: Holt, Rinehart & Winston, 1973.

MILLMAN, J., BISHOP, C. H., & EBEL, R. An analysis of test-wiseness. *Educational and Psychological Measurement*, 1965, 25, 707-726.

PRESTON, R. C. Ability of students to identify correct responses before reading. *Journal of Educational Research*, 1964, 58, 181-183.

PYRCZAK, F. Context-independence of items designed to measure the ability to derive the meanings of words from their context. *Educational and Psychological Measurement*, 1976, 36, 919-924.

SCIENCE RESEARCH ASSOCIATES, INC. *SRA achievement series answer keys, norms, and conversion tables, level C/forms 1 and 2.* Chicago, IL.: Author, 1979.

TUINMAN, J. J. Determining the passage dependency of comprehension questions in five major tests. *Reading Research Quarterly*, 1973-1974, 9, 206-223.

Accepted April 30, 1984.

APPENDIX E

Are Learning Disabled Students "Test-Wise?": An Inquiry into Reading Comprehension Test Items

Thomas E. Scruggs

Steve Lifson

Utah State University

Running head: READING COMPREHENSION TESTS

Abstract

Previous research has indicated that students in many cases can answer reading comprehension test questions correctly without having read the accompanying passage. The present research compared, in two experiments, the ability of learning disabled (LD) students and more typical age peers to answer such reading comprehension questions presented independently of reading passages. In Experiment 1, LD students scored appreciably lower under conditions resembling standardized administration procedures. In Experiment 2, reading decoding ability was controlled for; however, the performance differential remained the same. Results suggested a relative deficiency on the part of LD students with respect to reasoning strategies and test-taking skills. In addition, the validity of some tests of "reading comprehension" was discussed.

Are Learning Disabled Students "Test-Wise?": An Inquiry into Reading Comprehension Test Items

For many years, there has been some argument over what reading comprehension tests "really" measure (e.g., Thorndike, 1973-1974). The most commonly observed standardized reading comprehension item format consists of a passage and a number of associated multiple choice questions. Reading and understanding the passage is assumed to be a necessary pre-condition to correctly answering the questions. After examining the literature, however, one is forced to question the assumption of question dependence on the stimulus passage. Preston (1964) found that college students were able to answer reading comprehension items with the passages blacked out at a rate significantly above chance. Tuinman (1973-1974) administered five major tests to 9,451 elementary-level students under several conditions. Students in the no passage condition (relevant passage had been blacked out) on the average achieved only 30% fewer correct answers than subjects in the passage-in condition. Similar results were obtained by Pyrczak (1972, 1974, 1975, 1976) and Bickley, Weaver, and Ford (1968). A follow-up study of passage independence by Lifson, Scruggs, and Bennion (1984) revealed that passage-independent items are still quite common in elementary level achievement tests. College undergraduates were able to answer 75%, or almost 12 of 16 questions on the Stanford

Achievement Test, Level P-3, without reading the associated passages. This score is considerably above that expected by chance responding.

Scruggs, Bennion, and Lifson (1985) interviewed elementary age students regarding their responses on a reading comprehension test. They found that students often chose their answers based upon their own prior knowledge, rather than content of the reading passage. When students reported using such prior information, they answered correctly in over 60% of the cases.

Reading comprehension items which are independent of the associated passage can be answered on the basis of the following: (a) general knowledge, (b) interrelatedness of the questions on a particular passage, and (c) faulty item construction, i.e., keyed option is twice as long or more precisely stated (Pyrczak, 1975). In the first two cases, the presence of enough information in the question stem to identify the topic is an important factor (e.g., "Which of the following statements in NOT true of penguins?"). Such a stem may render a question answerable in terms of information already available to the examinee and provide clues to the answers of related questions about the same passage that lack such information in the stem ("This passage is about: a) birds of South America, *b) birds of the Antarctic .... etc."). The cues which individuals apply to a testing situation to maximize their score correspond to Millman, Bishop, and Ebel's (1965) criteria of test-taking skills, or "test wiseness."

While test constructors may be able to point to high validity coefficients for their reading comprehension tests and subtests, an important question arises concerning whether all students are equally able to answer questions with the above mentioned characteristics without reading the passage. Are some groups of students at a relative advantage/disadvantage in ability to answer these questions without reading the passage? To answer this question, a group of students classified as learning disabled (LD) and a group of regular classroom students were administered a selection of multiple choice reading comprehension questions with the relevant passages removed. The conditions of this experiment were meant to resemble those of a normal testing situation--i.e., students were required to read the questions without assistance. This did not permit us to determine the extent to which any observed differences between the regular and LD students were due to reasoning or variations in general knowledge between the two groups or simply reflected a difference in reading ability. To address this issue, a second experiment was performed to see if similar differences could be found when word reading was controlled for.

Experiment 1

Method

Subjects and Materials

Subjects consisted of 67 regular classroom and resource room third grade students selected from several elementary schools in a western rural area. Of these subjects, 52 were regular classroom students and 15 were classified as LD by P.L. 94-142 and local criteria, which included a 40% discrepancy between actual and expected performance in two areas of academic functioning. The average grade equivalent of the total reading score of the non-LD students on the Comprehensive Test of Basic Skills (CTBS) was 3.4 (SD=.8), while the average CTBS total reading score for the LD students was 2.1 (SD=.5).

Fourteen multiple choice reading comprehension questions without the accompanying passages were selected for this task. Items were drawn from the Stanford Achievement Test, Level P-3, Form E (1982). Items had been chosen to represent questions thought by the author to be answerable in terms of: (a) the general knowledge of the test taker, and (b) the degree to which the interrelatedness of the items served as a cue to the answers. These items were taken from the Lifson et al. (1984) study, in which students' ability to answer these questions had been documented. The items were kept in clusters which belonged together in terms of association with a particular passage.

Procedure

Treatment was administered in regular instructional groupings. Materials were passed out and all students were told that they were about to take a reading test for which they would not be shown the accompanying reading passages, but that they should try their best to answer all questions. No time limit was imposed upon the task.

Results and Discussion

The regular classroom group answered correctly approximately 55% of the questions, for mean score of 7.8 (SD=1.96). This score was significantly above a chance score of 3.5 ($\underline{t}(102) = 11.27$, $p<.001$). In contrast, the LD students answered correctly only 35% of the questions, for a mean score of 4.9, only slightly higher than chance ($\underline{t}(28) = 1.77$, ns). The obtained score of the non-LD group was significantly higher than the LD group ($\underline{t}(65) = 4.91$, $p<.001$).

The present findings suggest that regular classroom students are able to recognize and make use of cues in testing situations in order to increase their scores, even when reading passages are deleted, and "reading comprehension" supposedly cannot be measured. Apparently, LD students are not able to benefit equally from these cues. Since neither group should have scored above chance on a reading comprehension test with the reading passages deleted, it is possible that a certain amount of bias exists against children with learning disabilities on some standardized tests of reading comprehension. Students in regular classes when unable to read or otherwise obtain meaning from reading passages are still able to answer correctly comprehension questions.

Students with learning disabilities, however, do not seem to have these skills, and are thereby punished twice for a reading handicap: Once for being less able to read and comprehend the passage, and a second time for being unable to "second guess" test questions, as their nonhandicapped peers are apparently able to do.

One possible explanation for this discrepancy between LD and regular classroom students is that LD students are simply less able to read (decode) the questions, and for that reason are less able to outguess the test. That is, LD students are less deficient in "test taking skills" than they are in reading ability. In order to address this question, a second experiment was designed, in which ability to read would be controlled for. Although the conditions in this experiment could not parallel those of standardized test procedures, they did allow for an assessment of the extent to which differential scores are attributable to generally lower reading skills.

Experiment 2

Method

Subjects and Materials

The 42 subjects who participated in this investigation were different students drawn from the same population as those of Experiment 1, and consisted of 27 regular classroom third grade students and 15 third grade children classified as LD by P.L.

94-142 and local district criteria. Mean grade equivalent for the non-LD group (CTBS total reading) was 3.6 (SD=.9), and 1.9 (SD=.4) for the LD group. Materials were 14 items drawn from the Stanford Achievement Test, level P3, Form F, and were chosen on the same basis as those used in Experiment 1. Pages of the test were again left intact with questions left in the original order and the passages themselves blacked out during the copying process.

Procedure

Students were informed by their teacher that they were about to take a reading test without reading the corresponding passages. They were told to listen while the teacher read each item, and then answer the items. All students were given sufficient time to answer all questions.

Results and Discussion

The students in regular classrooms answered correctly 65% of the fourteen items, for a mean score of 9.14 (SD=1.8). The LD students, on the other hand, answered correctly only 45% of the items, for a mean score of 6.33 (SD=1.8). Although both obtained scores are well above chance, ($t(52) = 12.02$, and $t(28) = 4.325$, $ps<.001$, for regular classroom and LD students, respectively), the regular classroom group maintained its advantage over the LD students, $t(40)=4.87$, $p<.001$. The results suggest that learning disabled students may be less likely to apply test-taking strategies to reading comprehension questions to a degree of efficiency similar to their non-LD counterparts.

General Discussion

In Experiment 1, regular third grade classroom students were seen consistently to outscore their LD counterparts on a test of reading comprehension questions with corresponding passages deleted, and administered under conditions resembling standardized testing procedures. In Experiment 2, regular class third graders again outscored LD students, under conditions for which reading ability was controlled. The ability of third grade children in these cases to score 55% and 65% correctly on questions which refer to non-existent passages seems remarkable, and brings into question the issue of what some tests of "reading comprehension" are really measuring. Such passage independent items have been thought to assess test-taking skills and in fact have been used as measures of "test-wiseness" (e.g., Derby, 1978). Although it is suggested that differences in the use of test-taking strategies (such as use of prior knowledge, deductive reasoning, and elimination of implausable options) were responsible for much of the observed performance differences, other explanations are possible. Factors such as oral language decoding ability, attentional deficits, and test anxiety may have played a part in inhibiting performance on the part of the LD students. The role of these other factors in LD test performance is currently being investigated by the present authors (Scruggs, Bennion, & Lifson, 1984; Taylor & Scruggs, 1983). Whatever such tests are seen to measure, however, it is clear that: (a) it is not "reading comprehension," and (b) children classified as LD are at an apparent disadvantage.

An argument can be made that these comparisons are of trivial importance, since in standardized test administration, passages are not deleted; that all children in fact have equal access to passages which contain answers to reading comprehension questions. Although this argument has a certain face validity, some problems remain. First, since non-LD students can score so high on such items without reading the passages, the extent to which scores are a direct measure of "reading comprehension" seems uncertain. Second, since nearly all such tests are timed, students with incomplete understanding of relevant passages, but possessing an ability to "outguess" test questions under time constraints, clearly are at an advantage with respect to students not possessing such an ability. In this case, differences in scores on reading comprehension tests may in fact reflect in part a bias toward students with superior ability to respond to specific cues in the test-taking situation. As has been seen in the present experiments, LD students may well find themselves on the negative side of any such bias.

The extent to which LD and their non-LD counterparts differ on the present measures appears to have surprisingly little to do with reading ability. Although both groups gained when reading (decoding) ability was controlled for, each group was seen to exhibit the same degree of gain, amounting to about 10 percentage points for each group. Reported $\underline{t}$ values in Experiments 1 and 2 remained virtually identical. It seems clear, then, that much of the observed performance difference in Experiment 1 was due to skills other than reading ability, or "reading comprehension."

Two steps may be taken to help alleviate this potential source of bias. First, achievement tests should be revised so that reading comprehension tests directly assess comprehension of the provided passage. In fact, an informal review by the present authors of the major achievement tests indicates that many achievement test questions appear to be much less "passage independent" since the work of Tuinman (1973-1974) and others of a decade ago (Scruggs & Lifson, 1985). Second, it seems possible that at least some of these "test-taking skills" can be trained, and that this training may do much to correct this apparent disadvantage. The authors are at present investigating the effectiveness of such training (Taylor & Scruggs, 1983) and initial findings have been positive (Scruggs & Mastropieri, in press; Scruggs & Tolfa, 1985). Although such improved scores on tests may not necessarily reflect increased achievement, these scores could reflect more accurately achievement gains students have made, as evaluated by standardized achievement tests.

References

Bickley, A. C., Weaver, W. W., & Ford, F. G. (1968). Information removed from multiple-choice item responses by selected grammatical categories. Psychological Reports, 23, 613-614.

Derby, T. C. (1978). The effects of instruction in selected aspects of test-wiseness on the administration of standardized test items in the upper elementary schools. Unpublished doctoral dissertation, Southern Illinois University, Carbondale.

Lifson, S. A., Scruggs, T. E., & Bennion, K. E. (1984). Passage independence in reading achievement tests: A follow-up. Perceptual and Motor Skills, 58, 945-946.

Preston, R. C. (1964). Ability of students to identify correct responses before reading. Journal of Educational Research, 58, 181-183.

Pyrczak, F. (1972). Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. Reading Research Quarterly, 8, 62-71.

Pyrczak, F. (1974). Passage-dependence of items designed to measure the ability to identify the main ideas of paragraphs: Implications for validity. Educational and Psychological Measurement, 34, 34-348.

Pyrczak, F. (1975). Passage-dependence of reading comprehension questions: Examples. Journal of Reading, 19, 308-311.

Pyrczak, F. (1976). Context-indendence of items designed to measure the ability to derive the meanings of words from their context. Educational and Psychological Measurement, 36, 919-924.

Scruggs, T. E. Bennion, K. E., & Lifson, S. A. (1984, April). Spontaneously produced test-taking strategies of high and low comprehending elementary school children. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Scruggs, T. E., Bennion, K. E., & Lifson, S. A. (1985). An analysis of children's strategy use on reading achievement tests. Elementary School Journal, 85, 479-484.

Scruggs, T. E., & Lifson, S. A. (1985). Current conceptions of 'test-wiseness': Myths and realities. School Psychology Review, 14(3).

Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled students. Exceptional Children.

Scruggs, T. E., & Tolfa, D. (1985). Improving the test-taking skills of learning disabled students. Perceptual and Motor Skills, 60, 847-850.

Taylor, C., & Scruggs, T. E. (1983). Research in progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 50, 277.

Thorndike, R. L. (1973-1974). Reading as reasoning. Reading Research Quarterly, 9, 135-147.

Tuinman, J. J. (1973-1974). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.

Author Notes

The research described here was supported by a grant from the U.S. Department of Education, Special Education Programs, #G008300008. The authors would like to thank the excellent teachers of Cache Valley, Utah, for their assistance with this project: Marian Innocenti, Brenda Neiderhauser, Bonnie Olsen, Loila Anderson, and Edna Eams were particularly helpful. The authors would also like to thank Jill Barry, Ursula Pimentel, Marilyn Tinnakul, and Roberta LaMont for their assistance in the preparation of this manuscript. Address requests for reprints to Thomas E. Scruggs, Developmental Center for Handicapped Persons, UMC 68, Utah State University, Logan, Utah 84322.

APPENDIX F

LEARNING DISABLED STUDENTS' SPONTANEOUS USE OF TEST-TAKING SKILLS ON READING ACHIEVEMENT TESTS

Thomas E. Scruggs, Karla Bennion, and Steve Lifson

Abstract. **The present investigation was undertaken to identify tye type of strategies learning disabled (LD) students employ on standardized, group-administered achievement test items. Of particular interest was level of strategy effectiveness and possible difference sin strategy use between LD and nondisabled students. Students attending resource rooms and regular third-grade classes were administered items from reading achievement tests and interviewed concerning the strategies they had employed in answering the questions and their level of confidence in each answer. Results indicated that (a) LD students were less likely to report use of appropriate strategies on inferential questions, (b) LD students were less likely to attend carefully to specific format demands, and (c) LD students reported inappropriately high levels of confidence.**

Since the seminal article by Millman, Bishop, and Ebel in 1965, attention has been focused on test-taking skills, or test-wiseness, as a source of measurement error in group-administered achievement tests (Sarnacki, 1979). Defined as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score" (Millman et al., 1965, p. 707), test-wiseness is said to include such diverse components as guessing, time-use, and deductive reasoning strategies. Given that the effective use of such strategies may have little relationship to a particular academic content area, individuals or groups of individuals lacking in these skills may be at a disadvantage. A recently completed meta-analysis, for example, suggested that under certain circumstances, low-SES students are more likely to benefit from achievement test coaching than higher SES students — a finding which implies that low-SES students are relatively deficient in test-taking skills (Scruggs, Bennion, & White, 1984).

The present investigation was concerned with learning disabled (LD) children's spontaneous use of such strategies. Part of a larger investigation involving test-taking skills of exceptional students (Taylor & Scruggs, 1983), this study was conducted to identify possible deficits in test-taking skills on the part of LD children. Such deficits, if uncovered, would be helpful in developing remediation techniques.

Although much research has been conducted on nonhandicapped populations' test-taking skills (See Bangert-Drowns, Kulik, & Kulik, 1983; Sarnacki, 1979; and Scruggs, et. al., 1984, for reviews), little is known about LD students' test-taking skills. Scruggs and Lifson (1984) recently investigated LD students' differential ability to answer passage-independent reading-comprehension test items (i.e., reading-comprehension test items for which relevant passages had been omitted). Items were taken from standardized achievement tests known from previous research findings to be answerable by individuals who had not read the associated passage (Lifson & Scruggs, 1984), and thought

THOMAS E. SCRUGGS, Ph.D., is Research Evaluation Specialist, Exeptional Child Center, Utah State University.

KARLA BENNION, B.A., is a graduate student, Department of Psychology, Utah State University.

STEVEN A. LIFSON, M.Ed., is a graduate student, Department of Psychology, Utah State University.

BEST COPY AVAILABLE

BEST COPY AVAILABLE

to be a good measure of test-wiseness. In two experiements, nonhandicapped children scored 55% and 65% correct on such items, whereas students from the same grade scored much lower, even when word reading ability was controlled. Scruggs and Lifson (1984) argued that such findings also raised the question of what reading-comprehension tests do measure since no reading-comprehension test items should be answerable without prior reading of the associated passage. Scruggs and Lifson concluded that LD children may be at a relative disadvantage with respect to such test-taking skills as guessing, elimination, and deductive reasoning strategies applied to response items.

Scruggs, Bennion, and Lifson (in press) employed individual interview techniques to determine the nature of the strategies elementary-school children spontaneously produced on reading-achievement tests. Students representing a wide range of age and ability levels were given reading-achievement test items appropriate to their individual reading levels. Results indicated that students employed a wide range of strategies far beyond simply knowing or not knowing the answer, and that the use of these strategies was strongly predictive of performance. These findings provided valuable general information about the manner in which children respond to reading-achievement test items. However, the diversity of the population in age and achievement level was thought to have obscured observation of specific differences in test-taking skills between age or ability levels. The present investigation, therefore, was intended to determine whether LD and nondisabled students differed in strategy use on reading-achievement tests. In this investigation, grade level was held constant and the number of subtests was reduced to two: a reading-comprehension subtest, in which direct referring, elimination, and deductive reasoning strategies were thought to be important; and a letter-sound subtest, in which close attention to format demands was considered essential. In addition, since level of reported confidence was found to be a strong predictor of performance (Scruggs, et al., in press), and a prerequisite to strategy monitoring, confidence reports were examined for possible differences between ability groups.

Method

Subjects

Subjects were 32 third-grade students attending public schools in a Western university community. Twelve subjects were classified as learning disabled (LD) according to local school district criteria, which included a 40% discrepancy between ability and performance in two academic areas and PL 94-142 regulations. Twenty subjects were regular-class students, none of whom had been referred for special services or were considered by their teachers to be function at the highest achievement levels. Although the LD and regular-class students attended different schools, the schools were adjacent, drawing their populations from the same middle-class community. None of the students qualified for their schools' free lunch program. General cognitive ability appeared to be similar for the two groups. Mean Full-Scale IQ for the LD students (*Weschler Intelligence Scale for Childre-Revised*) was 92.75 (SD = 5.7). Mean Cognitive Skills Index for the non-LD students (*Test of Cognitive Skills*) was 96.16 (SD = 9.5)[1] Mean grade equivalent for reading comprehensions on the *Comprehensive Test of Basic Skills* (CTBS) for non-LD subjects was 3.9 (SD = .89), equivalent to a percentile score of 61. For LD studnets the mean CTBS reading-comprehension grade equivalent was 2.3 (SD = .29), equivalent to a percentile score of 21. The 16 boys and 16 girls constituting the sample were all 8-9 years old and Caucasian. Sex was evenly represented both in LD and non-LD groups.

Materials

Two reading tests were constructed from items taken from the *Stanford Achievement Test*. Test items were drawn from the Primary 2 battery for the instrument used with the LD group, whereas the Intermediate 1 level served as the source for the regular classroom group. Each test contained three reading passages with 14 dependent questions (10 content, 4 inference) on each form. Comprehension questions were left in their original order in relation to the selected passage. Questions were renumbered to avoid gaps where passages did not follow the sequential order of the original test. In addition, three items from the letter-sound test (level P3) were selected. These consisted of a stimulus word in which a letter or letters were underlined to represent a sound that the student was to identify among three options given below the stimulus word. These items served as distractors that closely matched the initial consonants of the stimulus word. For example, in the item:

blind
0 blink
0 nibble
0 leaned

leaned is the correct answer, since it contains the same sound as the underlined *l d* in the stem; *blink* is the distractor, containing the same initial consonant blend.

Procedure

Subjects, seen individually by one of two examiners, were asked to read the passages and questions aloud and mark the answers they thought were correct. Students were then told that they would be asked to state if they were sure/not sure that the selected answer was correct, and the manner in which they had chosen the particular answer. Subjects responses to the questions, "How did you choose that answer?" and "Are you sure or not sure of your answer?" were recorded verbatim on the protocol. Words the experimenters had previously deemed essential to answering the questions (key words) were marked in the examiner's copy of the instrument, and errors in these words were noted as the child read aloud.

Scoring

Test items were scored for correctness, confidence in answer (sure/not sure), and type of strategy reported. Two students from the non-LD group, who had misread more than 25% of the key words, were excluded from further analysis. The responses were divided into seven categories:

1 = Didn't know
2 = Guessed
3 = External source of knowledge (e.g., "I know all fish have scales")
4 = Referred to passage (e.g., "I read it")
5 = Quoted directly (e.g., "It says here that...")
6 = Eliminated options known to be incorrect
7 = Other reasoning (e.g., "It said comforted in the story. That sort of means relieved.")

Each response was evaluated in terms of the seven categories. Percent of agreement for scoring was assessed at 100% after each examiner scored 25% of the other examiner's protocols.

RESULTS

Results of *t*-test applied to percent of key words read incorrectly indicated that the groups did not differ significantly with respect to reading difficulty, $t(29) = .37$, $p > .20$. Overall, LD students misread 6.6% of 30 total key words, whereas non-LD students misread 6.75% of 29 key words.

Proportion correct by collapsed strategy group (inappropriate = strategies 1-3; referring = strategies 4-5; reasoning = strategies 1-3; referring = strategies 4-5; reasoning = strategies 6-7 was computed for item type and student group (see Figures 1 and 2).

Strategy data were scored for appropriateness of reported strategy. Strategies were considered appropriate if students reported referring to the passage on a recall question (strategy 4 or 5), or if they reported a reasoning strategy in response to an inferential question (strategy 6 or 7). Proportion of appropriate responses was then entered into a 2 group (LD vs. non-LD) by 2 item type (direct recall or inferential) analysis of variance (ANOVA) with repeated measures on the item-type variable. Because of the unequal group frequencies, a least-squares method of analysis (Winer, 1971) was employed. Significant differences were found for item type, $F(,29) = 9.19$, $p < .01$, and interaction, $F(1,27) = 7.58$, $p < .05$. Figure 3 depicts graphically the interaction effect. Although both LD and non-LD students reported a high proportion of referring to text strategies on recall questions (89% vs. 77%, respectively). Nonsignificant differences were observed for overall group means, $F(1,29) = 1.54$.

Analysis of confidence reports revealed that both groups were similar with respect to reported confidence level on *referring to passage* strategies with LD students reporting confidence in 85% of the cases and non-LD students reporting confidence in 92% of the instances. These reports were similar to actual performance, with correct scores of 81% and 86% on these items for LD and non-LD groups, respectively. On reasoning strategies, however, a different picture emerged. Here regular-class students were correct on 83% of the inferential items, compared to an average reported confidence of 71% of the items. The LD students, on the other hand, reported being confident an average of 95% of the cases, while being correct in only 63% of these cases.

Items on the letter-sound subtest were scored for responses which suggested attention to an inappropriate distractor. This inappropriate distractor took the form of an initial consonant blend present in the stem, but not underlined. A comparison of the number of inappropriate distractors by a group revealed significant differences, $t(28) = 2.47$, $p < .05$. Thus, LD students chose the inappropriate distractor in 52% of the cases, compared to the non-LD children who selected the inappropriate distractor in only 24% of the cases.

DISCUSSION

The present sample of LD third graders, with reading ability controlled for, differed from their regular-class counterparts with respect to (a) proportion of appropriate reasoning strategies reported for inferential comprehension questions, (b) performance and confidence level for items in which reasoning strategies had been reported, and (c) choice of an inappropriate distractor on a letter-sound test. However, LD students did not differ from their nondisabled peers in terms of appropriate strategy use on recall items. Generally, this sample of LD children was seen (a) to report fewer reasoning strategies, when appropriate, on reading comprehension-test items that their regular-class counterparts, and (b) to be less successful on those items for which they reported using reasoning strategies. These results support those reported by Scruggs and Lifson (1984) who found that LD students exhibited relatively inferior performance on a test of selected reading-comprehension test items for which the relevant passages had been removed, and for which reasoning strategies were thought to be necessary in order to answer the items correctly. The present finding of inappropriately high confidence levels exhibited by the LD students on items for which reaosning strategies had been applied supports a theory of a developmental deficit in meta-cognitive abilities (e.f., Torgesen, 1977), as inappropriately high confidence levels in task performance are often seen in younger children. Such a deficit on the part of LD children is thought to be critical, since ability to evaluate accurately a chosen response is a necessary prerequisite for effective test-taking.

LD students, tendency to attend to an inappropriate distractor may be a function of an attentional deficit (Krupski, 1980) on test format as much as a deficit in phonetic skills. It is unclear whether these test-taking skills are subject to remediation (Taylor & Scruggs, 1983), regardless, they may reflect a source of measurement error (Millman et al., 1965).

Reading comprehension seems to resist precise analysis and to be the subject of many theoretical orientations exist (Spiro, Bruce, & Brewer, 1980). If recall and inference are looked upon as two component parts of reading comprehension, however, results of the present investigation suggest that LD children demonstrate strategy and performance deficits on inference questions, but not on recall questions, with reading ability controlled for. Thus, it may be argued that the specific deficits exhibited here reflect problems in reading comprehension rather than test-taking skills. It seems likely, therefore, that strategy training in such areas could lead to improved reading comprehension as well as improved test-taking skills, particularly since selecting and implementing appropriate strategies has been found to improve general cognitive functioning (e.g., Torgesen & Kail, 1980). In the word-study skills subtest, however, the LD students apparently became confused by specific format demands which likely had little to do with the content being tested, (i.e., matching on initial consonant blend rather than an underlined vowel sound). Training for this type of strategy deficit, therefore, cannot be expected to bring about a concomitant increase in phonetic analysis skills.

Replication is necessary to further support and refine these findings. The present results suggested that LD children may benefit from specific training in (a) attending to specific format demands, (b) identifying inference questions, and (c) selecting and applying appropriate strategies relevant to such questions.

REFERENCES

Bangert-Drowns, R. L. Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. *Review of Educational Research*, 53, 571-585.

BEST COPY AVAILABLE

Krupski, A. (1980). Attention processes: Research, theory, and implications for special education. In B. Keogh (Ed.), *Advances in special education* (Vol. 1) (pp. 101-140). Greenwich CT: Jai Press.

Lifson, S., & Scruggs, T. E. (1984). Passage independence in reading comprehension items: A follow-up. *Perceptual and Motor Skills, 58*, 945-946.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness. *Educational and Psychological Measurement*, 25, 707-726.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, 49, 252-279.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press). An analysis of children's strategy use on reading achievement tests. *Elementary School Journal*.

Scruggs, T. E., Bennion, K., & White, K. (1984). *The effects of coaching on achievement test scores in the elementary grades: A meta-analysis.* Unpublished manuscript, Utah State University.

Scruggs, T. E., & Lifson, S. (1984). *Are learning disabled students 'test-wise'? An inquiry into reading comprehension test items.* Unpublished manuscript, Utah State University.

Spiro, R. J., Bruce, B. C., & Brewer, W. F. (1980). *Theoretical issues in reading comprehension.* Hillsdale, NJ: Erlbaum.

Taylor, C., & Scruggs, T. E. (1983). Research in progress. Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. *Exceptional Children, 50*, 277.

Torgesen, J. K. (1977). Memorization processes in reading disabled children. *Journal of Educational Psychology*, 69, 571-578.

Torgesen, J. K., & Kail, R. V. (1980). Memory processes in exceptional children. In B. Keogh (Ed.), *Advances in special education* (Vol. 1) (pp. 55-99). CT: Jai Press.

Winer, B. J. (1971). *Statistical principles and experimental design* (2nd ed.). New York: McGraw-Hill.

FOOTNOTES

The authors would like to thank Mrs. Bonnie Olsen, Dr. Ted Williams, director, and the staff of the Edith Bowen School, particularly Dorothy Dobson, for their valuable assistance with this project. The authors would also like to thank Marilyn Tinnakul and Jill barry for typing the manuscript.

Requests for reprints should be addressed to: Thomas E. Scruggs, UMC 68, Utah State University, Logan, UT 84322.

Although the tests of mental ability were different for the two groups, the tests were both standardized on the same scale with a mean of 100.

APPENDIX G

Perceptual and Motor Skills, 1985, 60, 467-470. © Perceptual and Motor Skills 1985

ATTITUDES OF BEHAVIORALLY DISORDERED STUDENTS TOWARD TESTS[1]

THOMAS E. SCRUGGS, MARGO A. MASTROPIERI,
DEBRA TOLFA AND VESNA JENKINS

Utah State University

Summary.—In two studies, attitudes reported toward testing by behaviorally disordered students and their regular classroom counterparts were compared. In Study 1, 12 behaviorally disordered and 25 average fifth and sixth graders were given a survey regarding their attitude toward tests and the test-taking experience. Students classified as behaviorally disordered reported less positive attitudes toward tests than their more average peers; these attitude differences were more pronounced on items which reflected subjective attitudes toward the test-taking situation and aspirations about performance and less pronounced on evaluation of the value of tests. In Study 2, which employed a sample of 25 behaviorally disordered and 25 regular classroom students matched on age and sex and used a longer attitude measure, differences were not found. Taken together, these studies suggest that attitudes toward tests are inconsistent in the two populations and that some behaviorally disordered students may not differ so much in this regard as supposed.

Students classified as having behavioral disorders have often been said to exhibit deficiencies in academic performance as measured by standardized achievement tests (Motto & Wilkins, 1968; Stone & Rowley, 1964). Kauffman (1981) has reviewed several studies which examined the academic achievement characteristics of behaviorally disordered students and concluded that often the performance of these students falls far below their potential. Bases of these academic deficits are not completely understood, but it is commonly thought that behavioral disorders exhibited by this population have a negative effect on academic achievement. It is possible, however, that other factors also play a role in the generally lower functioning of behaviorally disordered students. One of these factors may be a possible difference in attitude toward the evaluation process, particularly as evidenced by achievement tests. Since no data document possible differences in attitudes toward tests and the test-taking situation, the present pilot investigation was intended to provide information on whether behaviorally disordered students may differ from their more average peers with respect to attitudes with which they approach the test-taking situation. Results of such an investigation would not be expected to indicate causal relations between attitudes and test performance but might be of value to researchers interested in differences in characteristic performance on achievement tests between behaviorally disordered and more average students.

---

[1] The research described here was supported in part by a grant from the Department of Education, Special Education Programs, No. G008300008. The authors thank Ms. Cathy Smith, Coordinator of Special Education, Hillview Elementary School, Salt Lake City, Utah, for her assistance with this project. Address requests for reprints to Thomas E. Scruggs, Ph.D., UMC 68, Utah State University, Logan, Utah 84322.

STUDY 1

Method

Subjects were 37 fifth and sixth grade students attending a public school in a western metropolitan community. Twelve of these students had been classified as behaviorally disordered, and 25 were more typical fifth and sixth graders attending regular classes in the same school. The principal criteria for identification as behaviorally disordered were average ability coupled with social or emotional functioning substantially different from that ordinarily shown by some other students and supported by teachers' and psychologists' observations and reports. Identification as behaviorally disordered occurred after less intensive educational and psychological interventions had not remediated the observed deficiencies. All 12 behaviorally disordered students were attending a self-contained class in the same school as the more average fifth and sixth graders. The two groups were evenly distributed with respect to grade; the sample of more average students contained 12 fifth and 13 sixth graders, while the behaviorally disordered sample contained 6 fifth and 6 sixth graders.

The 12-item Test Attitude Survey was constructed as part of a larger investigation involving the test-taking skills of learning disabled and behaviorally disordered students (Taylor & Scruggs, 1983) and contained such items as "taking [illegible] test bothers me," "it is important for me to do well on a test," and "tests are unfair." "Yes" or "no" responses indicating aggreement or disagreement with the associated statement were solicited for each statement. Internal consistency of this survey had been reported as .78 (Kuder-Richardson 20) on a previous administration to regular class elementary school students, indicating a moderate level of reliability for a survey of this nature. Students were given the survey during regular classes and wrote an answer to each question as the teacher read each item aloud. Students were given 1 point for a positive response (i.e., "yes" to a positive statement, or "no" to a negative statement) and 0 points for a negative response. Tests were scored by independent scorers unaware of group membership.

Results

The reliability of the survey for the present sample was .76 (KR-20), which was consistent with previous reports. Comparison of total scores for the two groups indicated that the average group of students had scored more positively than the behaviorally disordered group. The regular fifth and sixth graders reported 63% positive responses ($M = 7.6$, $SD = 1.8$), while the behaviorally disordered students reported 47% positive responses ($M = 5.6$, $SD = 2.4$), a statistically significant difference ($t_{35} = 2.80$, $p < .01$).

In a supplementary analysis, factor analysis of responses for the group as a whole yielded three factors with eigenvalues greater than 1.00, which accounted for 67.5% of total test variance. A principal components analysis, using Kaiser's criterion for factor limitation, 1's in the diagonal, and varimax rotation (SPSS, 1983) yielded factors of personal feelings about tests (e.g., "taking a test makes me upset"), personal importance of tests (e.g., "it is important for me to do well on a test"), and evaluation of the worth of tests (e.g., "tests are unfair"). Items which loaded most highly on each factor were compared between the two groups by means of t tests. The two groups again differed on the first factor, subjective feelings about tests ($t_{35} = 2.34$, $p <$

.025), and Factor 2, subjective importance of tests ($t_{38} = 2.46$, $p < .02$); the two groups did not differ with respect to the third factor, evaluation of the value of tests ($t_{38} = .84$, $p > .05$).

Discussion

Present results suggest that this sample of behaviorally disordered children differed from their peers in attitudes expressed toward tests and the test-taking situation. Although the two groups did not appear to be different with respect to evaluation of the role of tests, they did differ in their personal feelings about tests. These findings seem to suggest that, although the present sample of behaviorally disordered students appeared to appreciate the worth or importance of tests, they reported much less positive personal feelings about tests.

Several issues, however, can be raised which preclude drawing conclusions from the present findings. First, the sample of behaviorally disordered students is of insufficient size to permit generalizations to a larger population or further subdivision, e.g., by sex. Second, the attitude measure had too few items to draw firm conclusions regarding subtest performance. Study 2, then, was conducted to (a) confirm the present findings on a larger sample of behaviorally disordered students and (b) expand the attitude survey to contain more subtest items.

STUDY 2

Method

Subjects were 75 regular classroom students representing Grades 3 to 6 in a western metropolitan public school, and 25 students attending self-contained classes for students with behavioral disorders, Grades 3 to 6, in the same school. A different test attitude survey was constructed to include two subtests of items suggested by the factor analysis of Study 1: (a) items which reflected feeling about self in a testing situation (e.g., "I feel good when I take a test") and (b) items which reflected feelings about the value of tests themselves (e.g., "Tests help the teacher to see what we know"). This instrument had been piloted on a different sample of 55 elementary school students. Assessment of reliability gave a KR-20 of .74 for 22 items, and two subtests a and b, above correlated weakly with each other (.11). This low correlation suggested that separate aspects of testing attitudes were being assessed.

The 22-item measure was then administered to the sample of behaviorally disordered students and their peers in the students' regular classrooms. Items were read to the students by their teachers.

Results

Reliability (KR-20) of the attitude measure was .75. Reliability of the subtest of "personal feelings" items was .64, while reliability of the "value of tests" subtest was .59. Because the two groups differed in distribution of age and sex, 25 subjects were drawn from the peer group which were matched with the behaviorally disordered students on these variables. The resulting samples were virtually equivalent with respect to age (126.0 mo. vs 125.9 mo. for behaviorally disordered and regular class, respectively) and sex distribution (21 members of each group were boys).

Analysis of attitude responses indicated that groups did not differ with respect to total score, score on "personal feeling" items, or score on "value of tests" items ($|t| < 1.00$ in all cases). On total items, scores for behaviorally disordered and regular classroom students were, respectively, 16.5 ($SD = 4.4$), and 15.9 ($SD = 3.1$) out of a possible 22 positive responses. For "personal feelings" items, scores were, in the same order, 10.3 ($SD = 2.9$) and 9.8 ($SD = 1.9$) out of a possible 13 positive responses. For "value of tests" items, scores were 6.2 ($SD = 2.0$) and 6.1 ($SD = 1.6$) out of 9 possible positive responses. Although a further breakdown by sex might have been interesting, the small number of girls in each group would not permit this.

GENERAL DISCUSSION

In Study 1, a small sample of behaviorally disordered students reported less positive attitudes toward tests than did their regular class peers. These differences appeared to reflect differences in personal feelings regarding the testing situation rather than attitudes concerning the utility and value of tests in general, although the number of items was too small for conclusions to be drawn. In Study 2, a larger sample of behaviorally disordered and regular students matched on sex and age did not differ with respect to reported personal feelings about tests, attitudes concerning the value of tests, or total attitude. Although subjects reflected several different grade levels, attitudes by grade level could not be assessed due to the potential confounding of grade level by classroom.

One possible reason for the discrepancy between Studies 1 and 2 is that the subjects in Study 1 were not for one reason or another, representative of a larger population of behaviorally disordered students. Another possibility, and one worthy of further investigation, is that the discrepant findings reflect the fact that Study 2 was conducted during the beginning of the school year, when attitudes are commonly thought to be higher, while Study 1 was conducted at the end of the previous year after students had recently experienced testing. Further research is necessary to assess this hypothesis. At present, however, it may be concluded that some behaviorally disordered children might not differ so much from those in regular classrooms with respect to attitudes toward testing as might be thought.

REFERENCES

KAUFFMAN, J. M. *Characteristics of children's behavior disorders.* (2nd ed.) Columbus, OH: Merrill, 1981.
MOTTO, J. J., & WILKINS, G. S. Educational achievement of institutionally emotionally disturbed children. *Journal of Educational Research*, 1968, 61, 218-221.
SPSS, INC. *SPSSX user's guide.* Chicago, IL: Author, 1983.
STONE, F., & ROWLEY, V. N. Educational disability in emotionally disturbed children. *Exceptional Children*, 1964, 30, 423-426.
TAYLOR, C., & SCRUGGS, T. E. Research in progress: improving the test-taking skills of LD and BD elementary students. *Exceptional Children*, 1983, 50, 277.

Accepted January 17, 1985.

APPENDIX H

Format Changes in Reading Achievement Tests:
Implications for Learning Disabled Students

Debra Tolfa, Thomas E. Scruggs, and Karla Bennion
Utah State University

Running head: FORMAT CHANGES

Abstract

It has been seen that children's scores on reading achievement tests vary not only with knowledge of content but also with the differing formats of test items. Teachers working with learning disabled children or children with attention problems may wish to choose standardized tests with fewer rather than more format changes. The present study evaluated the number of format and direction changes, across tests and grade levels of the major elementary standardized reading achievement tests. The number of format changes varies from one change every 1.2 minutes on the Metropolitan Achievement Test Level E1 to one change every 21.3 minutes on the P1 level of the Stanford Achievement Test. Teachers may wish to take this evaluation into account when considering use of standardized reading achievement tests for their students.

Format Changes in Reading Achievement Tests.

Implications for Learning Disabled Students

The validity of group administered achievement tests for learning disabled and remedial reading students has been questioned (Benson & Crocker, 1979). A score on a science test, for example, should reflect the student's knowledge of the content area and not be dependent on reading ability. It is important, therefore, for the test maker to recognize bias related to such reading material and to remove that bias (Benson & Crocker, 1979). Another potential source of bias has been identified as test formats and format changes (Carcelli & White, 1981). In one study of reading achievement, children's responses to test items of the same content, presented in different formats, varied from 45% to 92% correct (White, Carcelli, & Taylor, 1981). Although standardization procedures can compensate in part for the influence of test formats, it is important that a student's score reflect, as accurately as possible, his/her knowledge of the content being tested.

Children in grades lower than the fourth have attained significantly lower test scores when the major format change of using a separate answer sheet is introduced (Cashen & Ramseyer, 1969; Harcourt, Brace, Jovanovich, 1973; and Ramseyer & Cashen, 1971). The skill of completing the separate answer sheet appears to be developmental in nature. While first and second graders do not spontaneously or after training use separate answer sheets efficiently (Ramseyer & Cashen, 1971), third graders have been successfully trained in the use of separate answer sheets (McKee, 1967).

Learning disabled children, children with attention problems, and children functioning below grade level may be even more adversely affected by format changes. Scruggs, Bennion, and Lifson (in press) in a study conducted with third grade learning disabled students, demonstrated that LD students were more easily confused and distracted by novel formats. These novel formats include the use of separate answer sheets. Most standardized tests begin use of separate answer sheets in fourth grade; the fifth grade LD student, functioning two years behind, may also experience difficulty with this task (Scruggs & Tolfa, 1985). Scruggs and Tolfa (1985) have demonstrated that fourth grade LD students do perform less accurately and with less speed on separate answer sheets than do their normally functioning peers.

Given the extent to which different formats inhibit correct responding, and the lesser ability of children at earlier developmental stages as well as the learning disabled student and poor reader to adjust to major format changes, teachers of such students may wish to consider using reading achievement tests with less frequent (rather than more frequent) format changes. Teachers will prefer to use tests on which a student's scores are affected more by knowledge of content, than the ability to adjust quickly to format changes.

Teachers, however, do not often have the opportunity to alter district decisions on which standardized tests are administered. In such situations, training may be beneficial. Scruggs and Mastropieri (in press) demonstrated that BD and LD students could be successfully trained in test taking skills involved with format changes. Scruggs and Mastropieri (in press) found that the more complicated the formats, the greater were the training gains. Since format has been shown to be a variable influencing test performance, the present investigation intended to compare the number of format changes, across grade levels, of the major standardized reading achievement tests. Levels from kindergarten to seventh grade were included.

Procedure

Reading subtests of the following standardized tests were analyzed for format changes: the Stanford Achievement Test (SAT) levels Primary 1, Primary 2, Primary 3, Intermediate 1, Intermediate 2; the California Achievement Tests (CAT) levels 10-17; the Metropolitan Achievement Tests (MAT) levels Primary 1, Primary 2, and Elementary and Intermediate; the Iowa Tests of Basic Skills (ITBS) levels 7-13; the Comprehensive Tests of Basic Skills (CTBS) levels A-G; and the SRA Achievement Series levels A-F.

A format change was defined as a variation in the number of options per item, a change from column to row or row to column, a change in either ~~stem~~ any part of the item itself or options from word to picture to passage to question to cloze item. Comparisons across tests and grade levels were made by dividing the time allowed by the number of formats in the test. For example, 20 minutes/4 formats means that in this case, there is a format change every 5 minutes. Interrater agreement was established at 100% by two raters discussing and recoding any independent disagreements in coding.

Results

Format information specific to each individual test is presented in Table 1. The standardized test with the least number of formats is the Metropolitan Achievement Test, which has an average of 3 formats across levels. The standardized test with the least number of format changes is the SRA, which has an average of 6 format changes. The SRA levels have one change every 13-16 minutes. The test with the greatest number of formats is the California Achievement Test and the Iowa Test of Basic Skills, both of which have an average of 8 formats. The standardized test with the greatest number of format changes is the Stanford Achievement Test. The SAT has an average of 18 format changes with level 12 showing 32 format changes, or a change every 2.6 minutes.

Insert Table 1 about here

The mean of the format changes across grade levels varies from one change every 6.1 minutes at grades 2-3 to one change every 12.75 minutes at grades K.

Discussion

Children's test scores vary not only with knowledge of content, but also with the differing formats of test items. Teachers of children with learning or attentional difficulties may wish to consider various options to help ensure all possible bias is eliminated from standardized tests. Teachers and school districts should consider using standardized tests with the lower numbers of format changes. When it is not possible to change tests administered, the teacher should provide practice and training with difficult formats. In addition, if a teacher suspects that students have difficulty adjusting to new formats, she or he may prefer to use a test which allows a reasonable amount of time before switching to a different format. The number of format changes on the major standardized reading achievement tests varies from 1 change every 1.2 minutes on the Metropolitan Achievement Test to 1 change every 21.3 minutes on the Stanford Achievement Test.

Although the teacher should always exhibit caution when interpreting test results, extra care can be taken when problems with format changes are suspected.

References

Benson, B., & Crocker, L., (1979). The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 39, 381-387.

Carcelli, L., & White, K. R., (1981). Effect of item format on students' math achievement test scores. Unpublished manuscript, Utah State University.

Cashen, V. M., & Ramseyer, G. C., (1969). The use of separate answer sheets by primary age children. Journal of Educational Measurement, 6, 155-158.

Harcourt, Brace, & Jovanovich; Test Department (1973, June). The effect of separate answer document use on achievement test performance of grade 3 and 4 pupils. NY: Harcourt, Brace, Jovanovich, Special Report No. 24.

McKee, L. E. (1967). Third grade students learn to use machine-scored answer sheets. The School Counselor, 15, 52-53.

Ramseyer, G. C., & Cashen, V. M., (1971). The effects of practice sessions on the use of separate answer sheets by first and second graders. Journal of Educational Measurement, 8, 177-181.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press). Spontaneously employed test-taking skills of learning disabled students. Learning Disabilities Quarterly.

Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disabled and learning disabled students. _Exceptional Children_.

Scruggs, T. E., & Tolfa, D. (1985). _Use of separate answer sheets by learning disabled students_. Unpublished data, Utah State University.

White, K. R., Carcelli, L., & Taylor, C., (1981). _Effect of item format on students' reading achievement test scores_. Unpublished manuscript, Utah State University.

Test References

California Achievement Tests, Levels 10-19, Form C. CTB/McGraw-Hill, Monterey, California, 1977.

Comprehensive Tests of Basic Skills, Levels A-G, Form U. CTB/McGraw-Hill, Monterey, California, 1981.

Iowa Tests of Basic Skills, Levels 7-14, Form 7. A. N. Hieronymus, E. F. Lindquist, H. D. Hoover, et al. Houghton Mifflin Co., The University of Iowa, 1980.

Metropolitan Achievement Tests, Levels P1-E1, Form JS. G. A. Prescott, I. H. Balow, T. P. Hogan, & R. C. Farr. Harcourt Brace Jovanovich, New York, 1978.

Stanford Achievement Test, Levels P1-I2, Form [illegible] E+F. E. F. Gardner, H. C. Rudman, B. Karlsen, & J. C. Merwin. Harcourt Brace Jovanovich, New York, 1982.

SRA Achievement Series, Levels A-D, Form 1. Science Research Associates, Chicago, 1978.

Table 1

Format Change Information

Test	Level	Grade	# Minutes	# Format	# Minutes/ Format Change	# Format Changes	# Minutes/ Format Change
CAT	10	K.0-K.9	116	7	16.7	7	16.6
	11	K.6-1.9	57	8	5.2	11	7.1
	12	1.6-2.9	69	9	5.8	12	7.7
	13	2.6-3.9	69	11	2.8	24	6.3
	14-19	3.6-7.9	45	5	7.5	6	9
Mean/				/8	/7.6	/12	/9.3
CTBS	A	K.0-K.9	53	5	8.8	6	10.6
	B	K.6-1.6	45	5	5.6	8	9
	C	1.0-1.9	65	6	7.2	9	10.8
	D	1.6-2.9	64	8	7.1	9	8
	E	2.6-3.9	70	8	7.8	9	8.8
	F	3.6-4.9	69	9	6.3	11	7.7
	G	4.6-6.9	60	9	5.5	11	6.7
Mean/				/7	/6.9	/9	/8.8
ITBS	7	1.7-2.6	68	10	3.3	18	6.8
	8	2.7-3.5	68	12	2.3	40	5.7
	9-14	3-7	57	3	14.3	4	19
Mean/				/8	/6.8	/17	/10.5
MAT	P1	1.5-2.4	45	3	15.0	3	15
	P2	2.5-3.4	40	2	3.3	12	20
	E1	3.5-4.9	40	3	1.2	33	13.3
	Int	5.0-6.9	40	3	2.4	17	13.3
Mean/				/3	/5.5	/16	/15.4
SAT	P1	1.5-2.9	85	4	21.3	4	21.3
	P2	2.5-3.9	90	8	6.0	15	11.25
	P3	3.5-4.9	80	9	6.7	12	8.9
	I1	4.5-5.9	85	8	3.1	27	10.6
	I2	5.5-7.9	85	8	2.6	32	10.6
Mean/				/7	/8	/18	/12.5
SRA	A	K.6-1.5	97	6	13.9	7	16.2
	B	1.6-2.5	115	7	16.4	7	16.4
	C	2.6-3.5	85	6	14.2	6	14.2
	D	3.6-4.5	48	3	16.0	3	16
	E	4.6-6.5	50	4	12.5	4	12.5
	F	6.6-8.1	50	4	12.5	4	12.5
Mean/				/5	/14.3	/5.2	/14.6

BEST COPY AVAILABLE

APPENDIX I

Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis

Thomas E. Scruggs, Karla Bennion, and Karl White
Exceptional Child Center
Utah State University

Running head: IMPROVING ACHIEVEMENT TEST SCORES

Abstract

Results of 24 studies which investigated the effects of training elementary school children in test-taking skills on standardized achievement tests were analyzed using meta-analysis techniques. In contrast to all previous reviewers, the results of this analysis suggest that training in test-taking skills has only a very small effect on students' scores on standardized achievement tests. Longer training programs are more effective, particularly for students in grades 1-3, and for students from low socioeconomic status background. Results from previous reviews of this body of literature are critiqued and explanations offered as to why the results of the present investigation are somewhat contradictory to previous reviewers' conclusions. Suggestions for further research are given.

Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis

Since the seminal work of Millman, Bishop, and Ebel (1965), much attention has been directed to the influence of test-taking skills, or "test-wiseness," on scores of achievement tests. Assumptions from the past have included that test-wiseness is a substantially separate variable not strongly correlated with intelligence (Diamond & Evans, 1972), that test-taking skills are alterable by training, and that these skills would transfer to higher scores on achievement tests (Ford, 1973; Fueyo, 1977; Sarnacki, 1979).

Training materials have been created (some of which are commercially available) to teach "test-taking skills" (e.g., Mini-Tests, 1979 and Test-Taking Skills Kit, 1980), and claims have been made that such training leads to increased test scores (e.g., Fueyo, 1977; Jones & Ligon, 1981; Samson, 1984). The rationale for such training programs stems from the common practice of utilizing results from achievement tests to assist in making decisions about educational placement, programming, and evaluation. To the degree that achievement tests are measuring test-taking skills rather than mastery of the content being tested (e.g., reading, math), decisions about placement, programming, and evaluation may be incorrect (see Ebel, 1965, for additional discussion). Promoters of teaching test-taking skills have claimed that students would obtain higher scores if deficiencies in test-taking skills were remediated, thus resulting in a more valid indicator of how well the student had mastered the content the test was designed to assess.

Although efforts to reduce measurement error in standardized achievement testing are commendable, several questions remain:

1. Although many people have concluded that test-taking skills training leads to increased test scores, is that position consistently supported empirically, and what is the magnitude of typically obtained effects?

2. Can the cost of typical test-taking training programs be justified in view of the magnitude of observed effects and the alternative uses of the same resource (i.e., is it cost-effective?)?

3. Are some types of training more valuable than others in increasing performance on achievement tests, and are some groups of children more likely than others to benefit from such training? The purpose of the present investigation was to integrate the results from previous research to answer the preceding questions as they pertain to standardized achievement tests with elementary school-aged children.

Review of Previous Work

Several reviewers have previously examined the effects of teaching test-taking skills (Bangert-Drowns, Kulik & Kulik, 1983;

Ford, 1973; Fueyo, 1977; Jones & Ligon, 1981; Sarnacki, 1979; Taylor, 1981). A summary of the characteristics and conclusions of these reviewers is shown in Table 1.

Insert Table 1 about here

All previous reviewers concluded that test-taking skills could be taught effectively and resulted in benefits for children (including higher achievement test scores). Unfortunately, except for Bangert-Drowns et al. (1983) and Taylor (1981), previous reviews failed to indicate the procedures or criteria for including research studies in their review, did not cite and critique prior reviews, and apparently only analyzed results of the primary research included in their review in terms of the original researcher's conclusions. As will be shown below, all of the reviewers failed to include a substantial number of studies with elementary aged children. Consequently, one cannot be confident that results cited in these reviews are representative of available reseach. It is also difficult to draw conclusions about the magnitude of the alleged effect of training students in test-taking skills since most of the reviewers stated only that differences were found, or improvement was noted, and occasionally referred to statistically significant differences between groups. Without knowing more about the magnitude of the effect attributable to teaching test-taking skills, it is difficult to draw conclusions about whether it is likely to be a wise investment to divert resources from other activities (e.g., teaching reading) to teach test-taking skills.

Taylor (1981) conducted an excellent review on the effects of practice, coaching, and reinforcement on test scores. This investigation focused upon all age levels and on group-administered as well as individually administered tests. The great majority of studies selected, in fact, concentrated on either IQ tests or non-elementary age populations; consequently, a substantial number of studies which investigated the effects of training achievement test-taking skills with elementary-aged children were not included in her review.

The most comprehensive analysis to date of the effect of teaching test-taking skills on achievement test scores was a meta-analysis recently completed by Bangert-Drowns et al. (1983). The effect of teaching test-taking skills for elementary-and secondary-aged children was analyzed by computing a standardized mean difference effect size for each study (Glass, 1977) to indicate the extent to which achievement test scores were altered by training. This was a substantial improvement from most earlier reviews which relied primarily on authors' conclusions or tests of statistical significance without indicating the magnitude of effects. Knowing the magnitude of improvement is very important so that practitioners can make judgments concerning whether the investment in training is cost-effective compared to what else could have been accomplished with that time. Bangert-Drowns et al. (1983) concluded that teaching test-taking skills raised standardized achievement test scores by .25 standard deviations--enough to raise the typical student from the 50th to the 60th percentile. They also concluded that length of training program was positively related to effect size; drill and practice was less effective than training in "broad cognitive skills;" and effectiveness of training was not affected by identifiable subject characteristics or other characteristics of the program.

Although Bangert-Drowns et al. provided valuable information, their study is limited by several factors. First, a number of studies have been done which were not included in their review. Secondly, although indicators of study quality were coded, there was no report of efforts to determine if there were differential effects for studies of high versus low quality. It may be, for example, that investigations of lower quality produce effect sizes which are substantially different (and also less credible) than studies of high quality.

Third, their decision to average all outcomes from a given study into one measure of effect size can be misleading. For example, Levine (1980) randomly assigned low SES and not low SES fifth graders to either test-taking training or control groups and collected data on students' scores on standardized reading achievement and an assessment of "test-wiseness". Four obvious effect sizes are possible: low SES experimental versus control for reading and test-wiseness; and not low SES experimental versus control for reading and test-wiseness. These four effect sizes range from .38 to 1.52 and average .90. To report only the average of all four is not only misleading, but irretrievably obscures important differences between types of subjects and types of outcome (e.g., in this study the effects for low SES subjects were much larger than "not low SES" subjects for both outcomes, and effects for test-wiseness were much larger than reading achievement for both groups).

Finally, in some instances Bangert-Drowns et al. appear to have used inappropriate computations for determining the effect size. For example, in the Romberg (1978) study, classrooms were randomly assigned to treatments, and class averages were used as the unit of analysis. While the use of classroom means as the unit of analysis is an appropriate statistical procedure (Peckham, Glass, & Hopkins, 1969), the standard deviation of group means will generally be much smaller than the within-group standard deviation. The use of the between-class standard deviation will result in a much larger effect size and will not be comparable to studies for which the within-group standard deviation was used. In the Romberg study, Bangert-Drowns et al. apparently used the between-class standard deviation for achievement test scores and obtained an effect size of .48. By contrast, the present authors estimated the effect size (since within-group standard deviations were not reported) by converting the reported percentile scores to Z scores and using differences in Z scores as the effect size. This procedure yielded an effect size based on the within-group standard deviation of only .14--less than one third the magnitude of Bangert-Drowns et al. estimate.

Other important questions remain unaddressed by Bangert-Drowns et al. (1983). First, many investigations believe that the training of test-taking skills is particularly beneficial for children in low socioeconomic settings (e.g., Jones & Ligon, 1981; Jongsma & Warshauer, 1975). Thus, it is important to determine whether teaching test-taking skills has a differential effect on children of low socioeconomic status than it does on children who do not come from such groups. Secondly, it is important to determine whether the effects of training in test-taking skills are different for children of different ages. In the Bangert-Drowns et al. study, students in grades 1 to 6 were combined into one category. Third, it is important to replicate their findings about length of training and type of training, and to determine whether there are any other important concomitant variables or interactions among variables not identified by Bangert-Drowns et al. Finally, it is important to know whether studies of adequate validity produce different effect sizes from studies of less than adequate validity, and whether there is a differential effect for different types of dependent measures (e.g., achievement tests, measures of test-wiseness, student attitude).

Procedure

Location of studies. Several procedures were used to find as many studies as possible which investigated the effect on group-adminstered standardized achievement test scores of teaching test-taking skills to elementary-aged school children. Studies which examined attempts to improve, for example, scores on individualized achievement tests or IQ tests were excluded from this analysis. Also excluded from analysis were studies which investigated the effects of training on achievement test performance of students of greater than 6th grade level.

Studies were located by first conducting a computer-assisted search of Dissertation Abstracts International, Psychological Abstracts, and Educational Resources Information Center (ERIC) data bases. Studies found in this way were examined to determine whether they contained references to other appropriate studies. Previous reviews of research on teaching test-taking skills (Bangert-Drowns et al., 1983; Ford, 1973; Fueyo, 1977; Jones & Ligon, 1981; Sarnacki, 1979; Taylor, 1981) were also examined for additional studies. Twenty-four experimental studies of the effects of teaching test-taking skills on achievement tests for students in grades 1 through 6 were located. This number is 70% greater than the greatest number of studies involving achievement tests for elementary school children found by any previous reviewer.

Coding. Each study was coded for 14 different variables which described the type of subjects with whom the research was conducted, the type of training provided, the experimental design used, and the type of outcome data collected. The specific variables coded are reported in Table 2 in the results section. Interrater consistency was established by having two independent reviewers code each article. Wherever disagreement occurred, differences were resolved by discussion.

To enable the comparison of all outcomes across all studies, an effect size for each relevant comparison was computed (Glass, McGaw, & Smith, 1981). Effect size was defined as the mean difference between two groups divided by the standard deviation of the control group. When means and standard deviations are not reported in a study, effect sizes can also be calculated from other statistics such as t and F. Basic conventions for determining which effect sizes to code, and methods of calculation when means and standard deviations were not available, are given in Casto, White, and Taylor (1983).

In addition, obtained effect sizes were adjusted using Hedges' (1981) formula for bias correction of the effect size estimator before analyses were done. Although the correction procedure was used for all results in the present study, the authors agree with Bangert-Drowns et al. that the overall difference in effect sizes due to this correction procedure was trivial (only 1 out of 65 effect sizes changed by more than .01 of an effect size).

Results and Discussion

The 24 investigations of the effect of teaching test-taking skills resulted in 65 effect sizes which were relatively evenly distributed among studies. The mean effect size for all comparisons including achievement tests, tests of test-wiseness, self-esteem, and anxiety, was .21, a figure which is consistent with that of Bangert-Drowns et al. but should be interpreted with caution since it is the average across different types of dependent measures, studies of differing quality, and students with different characteristics.

Table 2 shows the mean effect size for all levels of the different variables coded in the meta-analysis. As can be seen,

---

Insert Table 2 about here

---

the average effect size for studies with adequate validity is relatively close to that of studies with inadequate validity (.20 vs. .29). Although this suggests that it may not be necessary to account for quality of study in interpreting the impact of training students in test-taking skills, further examination of Table 2 shows that this is not the case. In particular, we note that the average of 44 effect sizes for achievement test scores from studies of adequate validity is .10, while the average of 6 effect sizes from adequate studies measuring "test-wiseness" is .71--almost 10 times as large. There are also no measures of test-wiseness or measures such as anxiety, self-esteem, and attitude towards the test, which come from studies with inadequate validity. Thus, the apparent equivalence in average effect sizes between studies of adequate validity and inadequate validity is largely attributable to the fact that outcomes other than achievement all come from studies of adequate validity and yield substantially higher effect sizes than measures of achievement.

The mean effect size for achievement test scores from studies of adequate validity is only .10 compared to an average of .29 for achievement test scores for studies with inadequate validity. This contrasts sharply with the findings of Bangert-Drowns et al. who reported an average effect size of .25. Part of the reason that Bangert-Drowns et al. found a higher average effect size may have been that they collapsed several different outcome measures from the study into one average effect size. As noted above, this can be misleading and prevents analyses of important issues.

Because there is such a dramatic difference in average effect size between studies with adequate validity and studies with inadequate validity, and between measures of achievement and other measures, the remaining analyses will focus primarily on effect sizes of achievement tests from studies with adequate validity.

The mean effect sizes for achievement test scores from studies with adequate validity for different levels of length of treatment, SES level, and grade level are shown in Table 3.

Insert Table 3 about here

As can be seen, there was considerable difference between interventions which were less than 4 hours and those which were 4 or more hours (.04 vs. .29). A similar finding was seen when results of achievement test scores were broken down by grade level. When treatments were administered to students in the primary grades (1-3), the average effect size on standardized achievement tests was only .01. From grades 4-6, however, the mean effect size for achievement tests was much higher, .20. The difference between students of differing socioeconomic background was very slight ($\overline{ES}$ = .14 vs. $\overline{ES}$ = .09), with a very small advantage for students from low socioeconomic backgrounds.

Even more interesting than the average effect size for different levels of these three variables are the interactions between the variables. As can be seen in Figure 1, for treatments involving less than 4 hours, students in the primary grades exhibited slightly negative effect sizes ($\overline{ES} = -.12$) while students from grades 4 through 6 had an average effect size of .19. For students receiving more than 4 hours of training, however, there is no difference--students in both grades 1-3 and 4-6 had an average effect size of .29. Although the mean effect size for students in grade 1-3 with 4 or more hours of treatment is based on only four studies, these data are provocative and require further investigation. More specifically, it appears that for older students, a short amount of training in test-taking skills may result in substantial improvement. However, for younger children, it takes much more training before there are observable benefits.

Figure 2 shows another interesting interaction between length of training and socioeconomic status. With less than 4 hours of treatment, neither "low SES" nor "not low SES" subjects benefited appreciably (average effect sizes are .05 and .08). With high levels of treatment, students from low socioeconomic backgrounds benefit more than twice as much as students who are not from low socioeconomic backgrounds (average effect size = .44 vs. .20). Again, this finding requires further replication before confident conclusions can be drawn, but it suggests that authors who have contended that training in test-taking skills is most important for students from low socioeconomic background (e.g., Jones & Ligon, 1981; Jongsma & Warshauer, 1975) may be correct.

Before drawing conclusions about the efficacy of training students in test-taking skills, it is important to comment briefly on the differences in average effect sizes between outcomes of achievement test scores ($\overline{ES}$ = .10), tests of test-wiseness ($\overline{ES}$ = .71), and measures of anxiety, self-esteem, and attitude towards tests ($\overline{ES}$ = .44). Admittedly, the measures other than scores on achievement tests are based on a very limited number of studies, so one should be cautious in drawing conclusions. However, from these data, it appears that tests of test-wiseness are more sensitive to training effects. One explanation for this much larger average effect size is that the training program is "teaching to the test." The fact that high scores on tests of test-wiseness are not necessarily related to higher achievement test scores suggests that the relation between test-wiseness and high scores on achievement tests is not very strong. It should be remembered that the primary argument for providing training in test-taking skills to students has always been related to the need to reduce measurement errors in the child's standardized test score. To the degree that that is happening, it has been assumed that test scores would go up. Although the fact that test scores are not going up appreciably is not proof that scores are not more accurate, it still leaves the burden of proof upon those who claim that training in test-taking skills is beneficial. Higher scores on tests of test-wiseness are not sufficient evidence for those benefits.

Conclusions

As noted earlier, this integrative review was designed to answer the following three questions:

1. To what degree is the popular position that training in test-taking skills is beneficial for children supported by empirical evidence?

2. Do the data about the effect of teaching test-taking skills justify the use of resources for this purpose as opposed to alternative uses of the same resource?

3. Are some types of training more effective or are some groups of children more likely to benefit from training in test-taking skills?

In response to the first question, the results of this review stand out in contrast with all previous reviewers of the effects of training in test-taking skills. The most credible evidence (results from high quality studies limited to scores on standardized achievement tests), at least as it pertains to elementary school-aged children, does not demonstrate a sizeable benefit for teaching test-taking skills. The reason for these different conclusions is partly attributable to the use of more systematic techniques than used by many of the previous reviewers to identify the magnitude of the effect and how that effect covaried with other variables. More importantly, a larger number of studies was identified and quality of study and type of outcome was accounted for.

Is training in test-taking skills cost effective? The answer is not clear-cut. Clearly, benefits of a tenth of a standard deviation are relatively small (less than one month worth of gain in reading for an average third grader), but they were obtained at relatively little cost. Even the longest training program lasted only 20 hours, and the majority of effect sizes came from studies in which training lasted less than 4 hours. The question also depends in part on whether one is talking about children in grades 1-3 or grades 4-6. These data suggest that for older children, a limited amount of training can have a discernible effect. For younger children, more training is necessary. Also, the fact that a few studies (unfortunately, it is a very limited number) suggest that training in test-taking skills has some positive impact on anxiety, self-esteem, and attitude towards tests should not be forgotten. However, before it is accepted as fact, more research needs to be done. It is clear that a comprehensive analysis of previous research on training test-taking skills suggests that the benefits are not nearly so great as has typically been concluded.

Data from the meta-analysis do suggest that training in test-taking skills is differentially effective for various subgroups of children. The interactions between length of treatment and grade level, and length of treatment and SES are particularly provocative and deserve further research. In general, the meta-analysis supports the conclusion of Bangert-Drowns et al. that longer training programs are more effective. As a general strategy, it also appears that training is more effective in the upper elementary grades than in the lower elementary grades. Whether or not a training package includes practice tests, reinforcement, or drill and practice does not seem to be an issue about which we have sufficient data to draw conclusions. More research is needed before we can decide what types of training are most effective.

Should training in test-taking skills be pursued? Hopefully, the results of this analysis will temper some of the unfounded enthusiasm in support of training children in test-taking skills. However, it would be unwise to conclude that training in test-taking skills is unwarranted or detrimental. Although the effects of such training are small, the investment is relatively cheap, and there is some evidence that for particular groups of children, training in test-taking skills can have substantial effects. Those tentative conclusions need further research, but indicate an area worth pursuing.

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. Review of Educational Research, 53, 571-585.

Casto, G., White, K. R., & Taylor, C. (1983). An Early Intervention Research Institute: Efficacy and cost studies in early intervention. Journal of the Division for Early Childhood, 7, 5-17.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 9(2), 145-150.

Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall.

Ford, V. A. (1973). Everything you wanted to know about test-wiseness. Princeton, NJ: Educational Testing Service. (ERIC Reproduction Service No. ED 093 912)

Fueyo, V., (1977). Training test-taking skills: A critical analysis. Psychology in the Schools, 14, 180-184.

Glass, G. V (1977). Primary, secondary, and meta-analysis of research. Educational Research, 5, 3-8.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). Integrating research studies: Meta-analysis of social research. Beverly Hills, Ca.: Sage Publications.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 5, 107-128.

Jones, P., & Ligon, G. D. (1981). Preparing students for standardized testing: A literature review. Austin, Tx: Austin Independent School District. (ERIC Document Reproduction Service No. 213 768)

Jongsma, E. A., & Warshauer, E. (1975). The effects of instruction in test-taking skills upon student performance on standardized achievement tests. New Orleans, LA: New Orleans University, Department of Elementary and Secondary Education. (ERIC Document Reproduction Service No. ED 114 408)

Levine, M. A. (1980). Training in test-wiseness on reading scores of low and middle SES pupils (Doctoral dissertation, Yeshiva University, 1979). Dissertation Abstracts International, 40, 6242A. (University Microfilms No. 80-12,678).

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness. Educational and Psychological Measurement, 25, 707-726.

Mini-tests (1979). New York: Educational Solutions, Inc.

Peckham, P. D., Glass, G. V, & Hopkins, K. D. (1969). The experimental unit in statistical analysis. The Journal of Special Education, 3(4).

Romberg, E. (1978). The effects of test-taking skills and attitudes on validity of standardized achievement test scores of inner-city children (Doctoral dissertation, University of Maryland, 1977). Dissertation Abstracts International, 39, 832A. (University Microfilms No. 73-12,646).

Samson, G. E. (1984). Effects of training in test-taking skills on achievement: A quantitative analysis. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49(2), 252-279.

Taylor, C. (1981). The effect of reinforcement and training on group standardized test behavior. Unpublished doctoral dissertation, Utah State University, Logan.

Test-taking skills kit (1980). Herndon, VA: Evaluation and Assessment Service, Inc.

META-ANALYSIS REFERENCES

Butler, D. D. (1983). An assessment of the effects of instruction and practice on the test-wiseness of fourth graders as measured by changes in standardized test scores (Doctoral dissertation, University of Wyoming, 1982). _Dissertation Abstracts International_, _43_ (7-A), 2322.

Callenbach, C. (1977). The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second-grade students. _Journal of Educational Measurement_, _14_, 335-341.

Costar, E. (1980). Scoring high in reading: The effectiveness of teaching achievement test-taking behaviors. _Elementary School Guidance and Counseling_, _15_, 157-159.

Crowe, D. E. (1982). The use of practice programs to improve test scores of elementary school students (Doctoral dissertation, University of South Carolina, 1981). _Dissertation Abstracts International_, _42_ (7-A), 3116.

Derby, T. L. (1979). The effects of instruction in selected aspects of test-wiseness on the administration of standardized reading items in the upper elementary school (Doctoral dissertation, University of Pennsylvania, 1979). _Dissertation Abstracts International_, _39_ (12-A), 7236.

Dillard, M., Warrior-Benjamin, J., & Perrin, D. W. (1977). Efficacy of test-wiseness on test anxiety and reading achievement among Black youth. _Psychological Reports_, _41_, 1135-1140.

Eakins, D. J., Green, D. S., & Bushnell, D. (1976). The effects of an instructional test-taking unit on achievement test scores. Journal of Educational Research, 70, 67-71.

Jongsma, E. A., & Warshauer, E. (1975). The effects of instruction in test-taking skills upon student performance on standardized achievement tests. New Orleans, LA: New Orleans University, Department of Elementary and Secondary Education. (ERIC Document Reproduction Service No. ED 114 408).

Kalechstein, P., Kalechstein, M., & Docter, R. (1981). The effects of instruction on test-taking skills in second-grade black children. Measurement and Evaluation in Guidance, 13(4), 198-202.

Lagana, J. L. (1979). The effects of incentive motivation and test-wiseness coaching on the standardized reading test scores of third-grade students (Doctoral dissertation, University of California at Los Angeles, 1978). Dissertation Abstracts International, 39, 4198-4199.

Levine, M. A. (1980). Training in test-wiseness on reading scores of low and middle SES pupils (Doctoral dissertation, Yeshiva University, 1979). Dissertation Abstracts International, 40, 6242A. (University Microfilms No. 80-12,678).

Luddeke, N. S. (1972). The effect of motivational programs on standardized achievement test performance of disadvantaged third graders at two levels of test difficulty (Doctoral dissertation, University of Cincinnati, 1971). Dissertation Abstracts International, 33 (6-A), 2820.

Romberg, E. (1978). The effects of test-taking skills and attitudes on validity of standardized achievement test scores of inner-city children (Doctoral dissertation, University of Maryland, 1977). Dissertation Abstracts International, 39, 832A. (University Microfilms No. 73-12,646).

Schuller, S. M. (1979). A large-scale assessment of an instructional program to improve test-wiseness in elementary school students. New York: Educational Solutions, Inc. (ERIC Document Reproduction Service No. ED 189 143).

Shisler, C. L. (1973). A study of test performance of first graders under three conditions of motivation (Doctoral dissertation, University of South Carolina, 1973). Dissertation Abstracts International, 1714-A.

Slaughter, B. A. (1976). An examination of the effects of teaching and practice in test-taking skills on student performance on a standardized achievement test (Doctoral dissertation, University of Pittsburgh, 1976). Dissertation Abstracts International, 37, 1505A. (University Microfilms No. 76-19,931).

Stephenson, P. C. (1976). Improving the learning disabled child's score on machine-scored tests. Journal of Learning Disabilities, 9(2), 17-19.

Suber, J. S. (1980). The effects of a practice test and days for administration on the demonstrated achievement level of low achieving third grade students in South Carolina (Doctoral dissertation, University of South Carolina, 1979). Dissertation Abstracts International, 40 (11-A), 5833.

Taylor, C. E., & White, K. R. (1983). The effect of reinforcement and training on group standardized test behavior. *Journal of Educational Measurement*, *19*(3), 199-209.

Thomas, R. J. (1977). The effects on three methods of test anxiety and the achievement test performance of elementary students: Providing test-taking information, test-wiseness training, and systematic desensitization (Doctoral dissertation, University of Wisconsin, Madison, 1976). *Dissertation Abstracts International*, *37* (9-A), 5717-5718.

Tinney, R. E. (1969). The effect of training in test-taking skills on the reading test scores of fifth grade children of high and low socioeconomic levels (Doctoral dissertation, University of Minnesota, 1968). *Dissertation Abstracts International*, *30*, 595A. (University Microfilms No. 69-11,505).

Van Hoose, W. (1969). The efficacy of counseling in the elementary school. Ohio State University. (ERIC Document Reproduction Service No. ED 033 394).

White, K. R., Taylor, C., Friedman, S., Bush, D. & Stewart, K. (1983). *An evaluation of training in standardized achievement test taking and administration: Final report of the 1981-82 Utah State Refinements to the ESEA Title I Evaluation and Reporting System*. Utah State University and Utah State Office of Education.

Yearby, M. E. (1976). The effects of instruction in test-taking skills on the standardized reading test scores for white and black third-grade children of high and low socioeconomic status (Doctoral dissertation, Indiana University, 1975). *Dissertation Abstracts International*, *36*, 4426A. (University Microfilms No. 75-23,438).

Table 1

Characteristics and Conclusions of Previous Reviewers of the Effect of Teaching Test-Taking Skills

Author/year	# of experimental studies cited	Methods for selecting studies specified?	Previous reviewers cited and critiqued	Outcomes of experimental studies cited in terms of	Conclusions about effectiveness of training test-taking skills	Variables cited which covary with effect of training	Type of studies included
Bangert-Drowns et al./1983	30	Yes	No	Standardized effect size	Effective $\overline{ES}$ = .25	Length of training program, type of training	Achievement tests; elementary and secondary level
Ford/1973	24	No	No	Conclusions	Effective	None	Achievement, IQ, and aptitude tests; preschool through adult
Fueyo/1977	19	No	No	Conclusions	Effective	None	Achievement, IQ, and aptitude tests; preschool through adult
Jones & Ligon/1981	5	No	No	Conclusions	Effective	Maintenance of effect Socioeconomic status	Achievement, IQ, and aptitude tests; preschool through adult
Sarnacki/1979	17	No	No	Conclusions	Effective	None	Achievement, IQ, and aptitude tests; preschool through adult
Taylor/1981	34	Yes	Yes	Standardized effect size	Effective $\overline{ES}$ = .62	Type of training, unit of administration, quality of study, type of test (achievement vs. IQ)	Achievement, IQ, and aptitude tests; preschool through adult

Table 2

Mean Effect Size for All Levels of All Coded Variables

		Adequate validity			Inadequate validity		
		$\overline{ES}$	SD_{ES}	N_{ES}	$\overline{ES}$	SD_{ES}	N_{ES}
All studies		.20	.40	55	.29	.33	10
Total sample size for study:	Small (0-75)	.32	.28	21	.40	.46	5
	Medium (76-150)	.11	.50	24			
	Large (150+)	.15	.30	10	.18	.08	5
Grade level:	1st-3rd	.03	.51	25	.14	.06	6
	4th-6th	.33	.39	30	.59	.54	3
Socioeconomic status level:	Low	.18	.37	37	.33	.36	8
	Not low	.24	.46	18	.11	.02	2
Use of reinforcement procedures as part of training:	No	.22	.40	48	-	-	-
	Yes	-.00	.43	7	.29	.33	10
Hours of training:	Less than 1 hr	.09	.43	14	.37	.47	5
	1 to 3 hrs	.09	.30	22	-	-	-
	4 hrs+	.40	.42	19	.20	.13	4
Use of practice tests as part of training:	No	.22	.43	42	.40	.46	5
	Yes	.12	.30	13	.16	.07	4
Ability level of students:	Mixed	.20	.52	47	.29	.33	10
	High ability	.09	.21	3	-	-	-
	Low ability	.31	.12	5	-	-	-
Type of assignment to groups:	Random	.27	.39	40	.30	.40	7
	Good matching	.24	.01	2			
	Poor matching	-.05	.37	13	.28	.10	3
Blinding of data collector:	Yes	.13	.44	34	.16	.07	4
	No	≈.31	.30	21	≈.38	.42	6
Type of outcome measure:							
Achievement test		.10	.33	44	.29	.33	10
Test-wiseness test		.71	.57	5	-	-	-
Other (anxiety, self-esteem, attitude)		.44	.36	6	-	-	-

$\overline{ES}$ = mean effect size for a particular group.

SD_{ES} = standard deviation of effect size distribution for a particular group.

N_{ES} = number of effect sizes on which a computation is based.

Note. Several other variables including Percent Male, Percent Handicapped, and Percent Minority were coded to determine whether mean effect size covaried with such subject characteristics. Results for those variables are not reported here because of infrequent reporting (e.g., Percent Handicapped could only be coded for 2% of the ES's), or lack of variance (e.g., 97% of the ES's for Percent Male fell between 47% and 54%).

Table 3

Mean Effect Sizes on Achievement Test Scores, Broken Down by Treatment Length, SES Level, and Grade Level

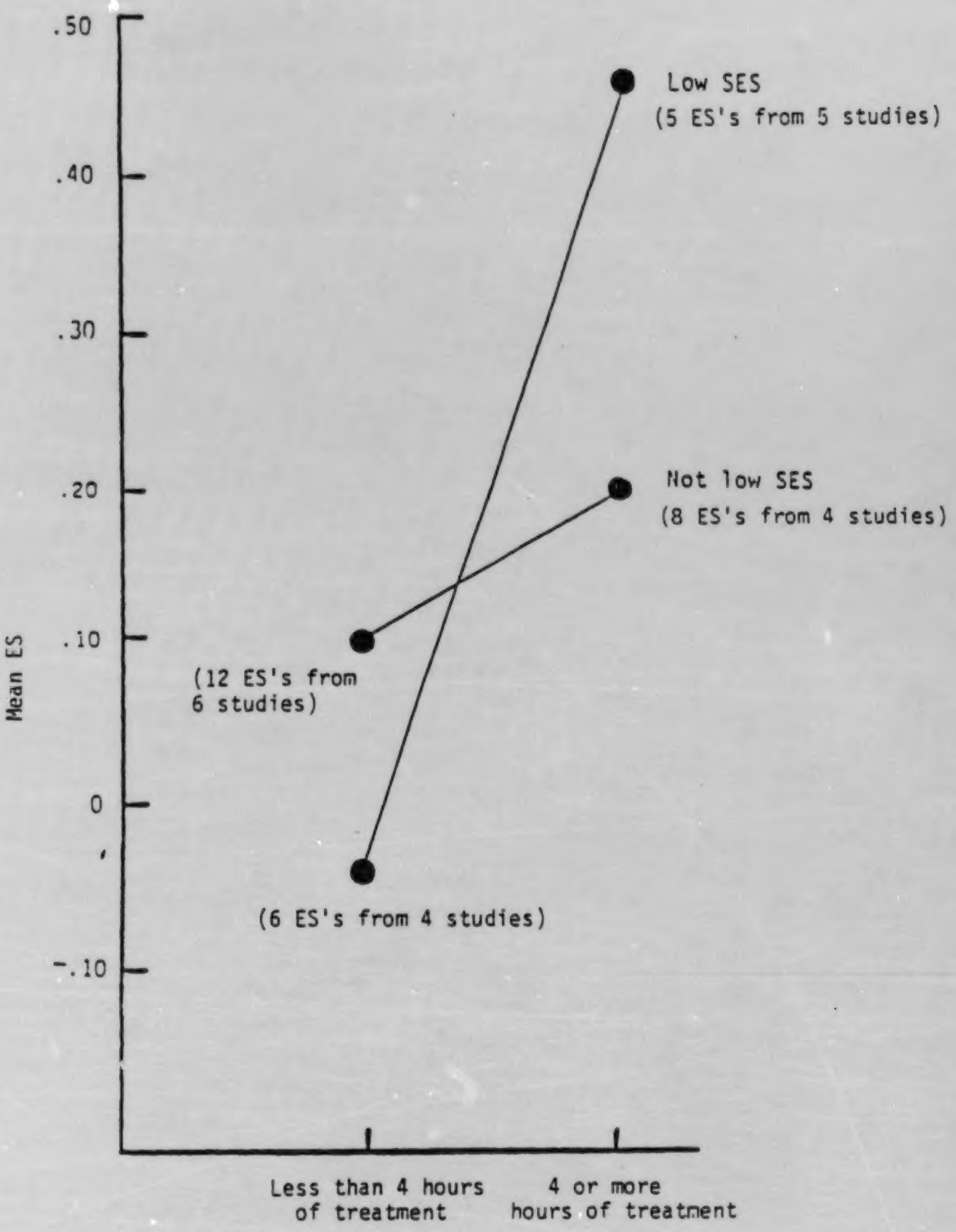
	Mean $\overline{ES}$	SD_{ES}	n_{ES}	$N_{Studies}$
Less than 4 hours of treatment	.04	.30	18	7
4 or more hours of treatment	.29	.31	13	8
Low SES	.14	.38	13	10
Not low SES	.09	.31	31	13
Grades 1-3	.01	.37	22	9
Grades 4-6	.20	.26	22	9

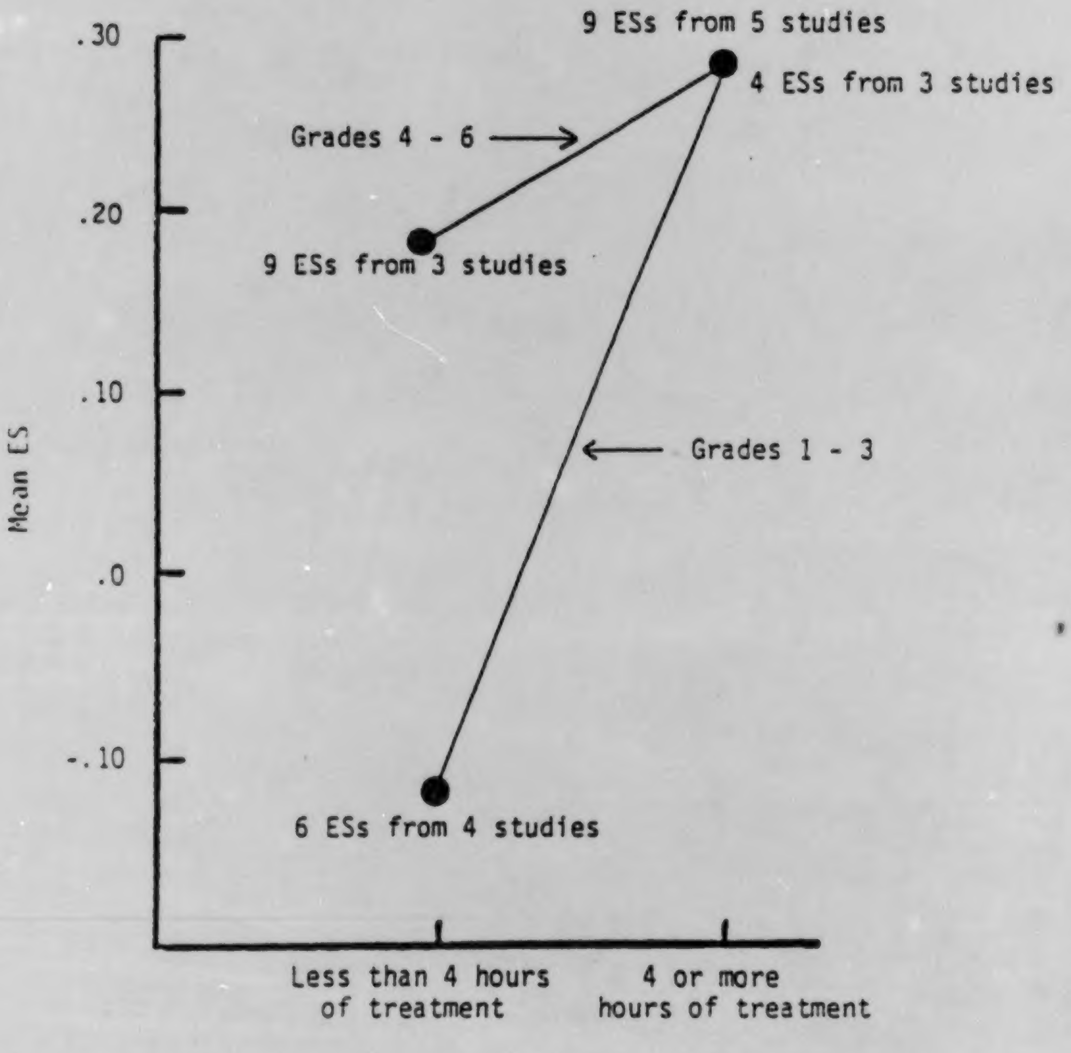
*Achievement test scores, studies with adequate validity only.

Figure Captions

Figure 1. Mean effect size by treatment length and grade level.

Figure 2. Mean effect size by treatment length and SES.


.50
.40
.30
.20
.10
0
-.10
Mean ES
Low SES
(5 ES's from 5 studies)
Not low SES
(8 ES's from 4 studies)
(12 ES's from 6 studies)
(6 ES's from 4 studies)
Less than 4 hours of treatment
4 or more hours of treatment


9 ESs from 5 studies
4 ESs from 3 studies
Grades 4 - 6
9 ESs from 3 studies
Grades 1 - 3
6 ESs from 4 studies
.30
.20
.10
.0
-.10
Mean ES
Less than 4 hours of treatment
4 or more hours of treatment

APPENDIX J

Perceptual and Motor Skills, 1985, 60, 847-850. © Perceptual and Motor Skills 1985

IMPROVING THE TEST-TAKING SKILLS OF LEARNING-DISABLED STUDENTS[1]

THOMAS E. SCRUGGS AND DEBRA TOLFA

Utah State University

Summary.—16 learning-disabled second- and third-grade students were matched on previous years' achievement scores and grade and assigned at random to experimental and control conditions. Students in the experimental condition were given 8 20-min. sessions of training in test-taking skills particular to the Stanford Achievement Test. Analysis of test scores indicated trained students scored significantly higher on one subtest of a shortened version of the test than students who had not been trained.

Since the seminal article by Millman, Bishop, and Ebel in 1965 on test-wiseness, or test-taking skills, interest has grown in the construct of test-wiseness as a possible source of measurement error (5). Although some specific groups and populations have been said to be low in "test-wiseness" (9), the issue of whether or not students classified as learning disabled exhibit the same test-taking skills as nondisabled peers has only recently been investigated (10). Scruggs and Lifson (7) administered reading comprehension test items with accompanying passages deleted to groups of learning-disabled and nondisabled students. Their results indicated that, although nondisabled students were able to take advantage of prior or partial knowledge and deductive reasoning strategies to answer most of the questions correctly, learning-disabled students were less able to utilize these strategies. In another investigation (6) learning-disabled and nondisabled students were interviewed regarding their strategies on reading-achievement-test items. Results suggested that learning-disabled students were less likely than their nondisabled peers to apply "appropriate test-taking strategies" to reading-comprehension-test items and learning-disabled students were more likely than nondisabled peers to be misled by particular format demands on tests of "word-study skills" (i.e., phonetic analysis).

Although the above research indicates that learning-disabled students may be lacking with respect to specific test-taking skills, this research does not indicate that these students can easily be taught these skills to the extent that achievement-test performance would improve. In fact, little is known about teaching test-taking skills to learning-disabled students. Recently, Dunn (2) successfully taught test-taking skills to a sample of junior high school-age learning-disabled students, but to date, test-taking skills have not been taught elementary-aged learning-disabled students. The purpose of the present research was to determine whether specific test-taking skills could be taught to elementary-aged learning-disabled students to improve their performance on standardized achievement-test items.

---

[1] This research was supported in part by a grant from the Department of Education, Office of Special Education, No. G008300008. The authors thank Marilyn Tinnakul and Mary Ellen Heiner for their assistance in the preparation of the manuscript. Address requests for reprints to Thomas E. Scruggs, Ph.D., UMC 68, Developmental Center for Handicapped Persons, Utah State University, Logan, Utah 84322.

METHOD

Subjects were 16 second- and third-grade learning-disabled students attending special education classes in a western metropolitan area.[1] Criteria for placement as learning disabled included average intelligence coupled with 40% discrepancy between ability and at least two areas of academic functioning. Although IQs were not available for this study, all students were said to have been functioning within a normal range of intelligence. Students were individually matched on the basis of grade and previous year's reading test scores and assigned at random to either experimental or control groups. Average reading percentile was 29.0 ($SD = 18.5$) for the experimental group and 28.3 ($SD = 19.7$) for the control group. Average age for each group was 7 yr., 8 mo. (*SDs* 8 mo. and 6.5 mo.; ranges 7 yr. to 8 yr., 4 mo. and 7 yr., 1 mo. to 8 yr., 6 mo., respectively, for experimental and control groups). Five (62.5%) second graders and three (37.5%) third graders were in each group; the experimental group contained four girls and four boys, while the control group contained three girls and five boys.

Materials were eight scripted lessons for each grade in a direct-instruction format and accompanying workbooks for students which included pencil-and-paper practice activities.[2] All items were similar to, but not exact items from, the Stanford Achievement Test. The general test-taking strategies taught in these materials included attending, marking answers carefully, choosing the best answer carefully, error-avoidance strategies, and appropriate situations for soliciting the teacher's attention. Specific test-taking strategies were taught for each reading subtest in the Stanford Achievement Test. These included structured practice on specific test formats for each subtest, and specific application of general test-taking strategies to each specific subtest. For example, with respect to the "letter-sound" component of the Word Study Skills subtest, students were taught to employ the following sequence of strategies: Look at and read the first word. Pronounce to yourself and think of the sound of the underlined letter. Carefully look at the underlined choices and choose the word with the same sound as the underlined letter. If you don't know all the words, read the words you do know or read parts of individual words you may know. If you're not sure of the answer, see if there are some answers that you are sure are not correct and eliminate those. Color in the answer quick, dark, and inside the line. Guess if you are still not sure; never skip an answer.

Experimental subjects were taught in small groups for four 20-min. lessons per week for 2 wk. Positive responding and attention to task were reinforced with stickers.

The first seven sessions taught the use of test-taking strategies within the specific context of each of the reading-related subtests. The last session consisted of a general review of all previous procedures. Each day of instruction involved extensive work with practice activities applied to practice test items. Students were given no information concerning the content of the actual test not specified in the published test directions.

---

[1] A small group of fourth-grade learning-disabled students was originally intended for inclusion in the study but had to be dropped because attrition and methodological problems were associated with the test administration for this group.

[2] T. E. Scruggs & J. Williams, SUPER SCORE: test-taking manuals and workbooks. (Unpublished training materials, Utah State University, 1984)

Following the last training procedure and posttest, all trained and control students were administered shortened versions of the reading subtests of the Stanford Achievement Test.

Items were taken from the Primary 2 level, Form E and Primary 3 level, Form E. The shortened version for Primary 2 level included the first 13 items on the Comprehension subtest and the first 16 items on the Word Study Skills subtest. The shortened version for the Primary 3 level included Items 9 to 22 on the Comprehension subtest and Items 1 to 9 and 19 to 32 on the Word Study Skills subtest. The Primary 2 test had a total of 13 Comprehension questions and 16 Word Study questions; while the Primary 3 test had a total of 14 Comprehension questions and 23 Word Study questions. The number of items was chosen for each condition to represent the number of items expected to be completed in 20 min., according to directions. Although the subtests were shortened to accommodate the student's scheduling constraints, standardization procedures were adhered to in the administration of the test, which was done in the resource setting by an administrator unfamiliar to the students and unaware of group membership of the students. Percent correct scores were analyzed instead of mean number correct because there was a different total number of items for each subtest and level.

RESULTS AND DISCUSSION

Percent correct scores for experimental and control students were compared statistically by means of *t* tests for independent means,[4] for Word Study Skills, Reading Comprehension, and combined subtests. Descriptively experimental students scored an average of 77.1% ($SD = 13.6$), 48.9% ($SD = 32.3$), and 63.0% ($SD = 20.6$), for Word Study Skills, Reading Comprehension, and combined subtests, respectively. Control students, by contrast, scored 56.8% ($SD = 20.1$), 50.3% ($SD = 24.3$), and 55.4% ($SD = 15.1$) on the same subtests. The only significant difference between groups was on the Word Study Skills subtest ($t_{14} = 2.38$, $p = .03$). Differences were not found on either the Reading Comprehension ($t_{14} = -.10$) or the total subtest ($t_{14} = 1.05$) scores.

It was seen that learning-disabled students trained in test-taking skills significantly outperformed their untrained peers on the Word Study Skills subtest but not the Reading Comprehension subtest, of a modified version of the Stanford Achievement Test. Although it is not certain why performance was improved on one subtest but not another, it is possible that performance on the Word Study Skills subtest was more easily trained because this subtest contained several different formats, introduced over a short period of time, which may have been confusing to the control students. The resulting effect size of this subtest (1.01 *SD* units) as well as the total score effect size (.63 *SD* units) are substantially larger than those reported in the literature (1, 8) and may indicate the deficit in test-taking skills may be somewhat stronger for this sample than others as supported by recent research.

---

[4] Since subjects were matched, it is possible to compute *t* tests for correlated data; this was not done here since scores of matched subjects were not correlated on the posttest. Corresponding *t* ratios for correlated data ($df = 7$) were essentially equivalent at 2.20 ($p = .06$), −0.10, and 1.12 for Word Study Skills, Reading Comprehension, and total subtests, respectively.

At least some aspects of the training appear to have been effective in raising test performance; however, the use of a no-treatment control group prohibits drawing conclusions regarding what specific aspects of the training were most effective. Further research could help clarify these variables.

Although it is true that the use of standardized achievement tests in special education is a controversial issue (4), it is also true that it is the obligation of special education personnel to maximize the functioning of learning-disabled students whenever possible, including performance on standardized achievement tests. It is also true that the skills taught for use on the Stanford Achievement Test may be even more valuable for teacher-made tests which may contain even more cues for the effective use of test-taking skills. Although the findings of the present investigation are promising, the small sample and the reduced version of the Stanford Achievement Test used as a dependent measure indicate that replication of these findings is necessary.

REFERENCES

1. BANGERT-DROWNS, R. L., KULIK, J. A., & KULIK, C. C. Effects of coaching programs on achievement test scores. *Review of Educational Research*, 1983, 53, 571-585.
2. DUNN, A. E., JR. An investigation of the effects of teaching test-taking skills to secondary learning disabled students in the Montgomery County (Maryland) Public School Learning Centers. Unpublished doctoral dissertation, George Washington Univer., 1981.
3. MILLMAN, J., BISHOP, C. H., & EBEL, R. An analysis of test-wiseness. *Educational and Psychological Measurement*, 1965, 25, 707-726.
4. SALVIA, J., & YSSELDYKE, J. E. *Assessment in remedial and special education.* (2nd ed.) Boston, MA: Houghton Mifflin, 1981.
5. SARNACKI, R. E. An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, 1979, 49, 252-279.
6. SCRUGGS, T. E., BENNION, K., & LIPSON, S. A. Spontaneously employed test-taking strategies of learning disabled students on reading achievement tests. *Learning Disability Quarterly*, in press.
7. SCRUGGS, T. E., & LIPSON, S. A. Are learning disabled students 'test-wise?' An inquiry into reading comprehension test items. Paper presented at the annual meeting of the American Education Research Association, Chicago, 1985.
8. SCRUGGS, T. E., WHITE, K., & BENNION, K. Teaching test-taking skills to elementary-age students: a meta-analysis. In T. E. Scruggs, Administration and interpretation of standardized achievement tests with learning disabled and behaviorally disabled elementary school children. Final Report, Grant No. G008300008, Department of Education, Special Education Programs, Washington, DC, 1984.
9. SLAKTER, M. J., KOEHLER, R. A., & HAMPTON, S. A. Learning test-wiseness by programmed texts. *Journal of Educational Measurement*, 1970, 7, 247-254.
10. TAYLOR, C., & SCRUGGS, T. E. Research in progress: improving the test-taking skills of LD and BD elementary students. *Exceptional Children*, 1983, 50, 277.

Accepted March 25, 1985.

APPENDIX K

The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children

Thomas E. Scruggs, Karla Bennion, and Joanne Williams

Utah State University

Running head: EFFECTS OF TRAINING

Abstract

Fifty-eight third graders from two elementary school classrooms were assigned at random to test-training and placebo groups. Students in the test-training group received six sessions of test-wiseness training specifically tailored to the Comprehensive Test of Basic Skills. Students in the placebo group received six sessions of creative writing exercises. The effectiveness of this training on achievement test scores was obscured due to the presence of ceiling effects. Supplementary analyses, however, provided some limited support for the effectiveness of this training. Trained and untrained groups were not seen to differ on measures of on-task behavior during the testing situation. An analysis of reported attitudes toward tests taken immediately after the three-day testing period suggested that (a) the standardized test experience was a stressful one for control subjects, and (b) that the test-wiseness training had exerted a significant ameliorating effect on attitudes in the treatment group. Results suggested that test-wiseness training may reduce levels of anxiety in elementary school children during test situations.

The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children

In recent years, the effectiveness of coaching on achievement test performance has been well studied (see Sarnacki, 1979, and Fueyo, 1976, for reviews). In a recent meta-analysis, Bangert-Drowns, Kulik, and Kulik (1983) determined that coaching for achievement tests in the elementary grades produced a generally facilitative effect (average effect size = .29) over all studies reviewed. More recently, Scruggs, Bennion, and White (1984) have argued that although training in test-taking skills does often produce an effect in the elementary school grades, this effect is dependent upon other factors, for example, length of training, age of students, and economic level of the students trained. Although researchers in the area of test-wiseness training have often examined variables in addition to actual test scores such as performance on test-wiseness tests and self-esteem, they have not addressed the issue of whether or not such training changes in any way the attitudes of elementary school children toward tests. This in itself could be an important finding for, concerning the degree to which school-age children are subjected to testing procedures, it would be helpful to ensure that such tests were not unnecessarily stressful. In addition, whether or not training in test-taking skills has a facilitative influence on the level of effort the students put into the test situation remains unclear. Such effort may be evaluated by means of the amount of time on-task students exhibit during standardized testing.

The present investigation was intended to address some of these issues by providing training in test-taking skills to a sample of third grade students and assessing, in addition to test performance, reported attitudes towards the test-taking experience and percent of time actually spent on-task during test administration. Although the effects of test-wiseness training have been well-documented in the past, the present investigation was intended to shed some light on peripheral issues and to address more specifically exactly what changes in attention and attitude occur as a result of coaching on achievement tests.

Method

Subjects

Subjects were 58 elementary-age school children attending the third grade in two different classrooms at a western rural school district. Sex was evenly distributed. Subjects were selected at random from both classes to participate in treatment and placebo groups.

Materials

Materials included a manual with six scripted 20- to 30-minute lessons in test-taking skills specifically tailored to the reading subtests of the Comprehensive Test of Basic Skills (CTBS), Level E. These materials were developed specifically for this project and included student workbooks for practice activities by the students (Williams, 1984).

Procedure

Over a two-week period, treatment students were administered six lessons in test-taking skills appropriate to the reading subtest of the CTBS, by a trained, outside experimenter. These lessons included, for example, time-using strategies, deductive reading strategies, error avoidance strategies, and specific practice activities in each of the subtests. To control for possible Hawthorne effects, the placebo group was given six exercises in creative writing by an outside experimenter at the same time treatment students were receiving test training. Within three days after the conclusion of training, students were given the CTBS by their regular classroom teachers in their regular instructional classes. During the taking of this test, observational measures were taken of on-task behavior of students by four trained observers unaware of group memberships of the students being observed. The observers employed a time-sampling procedure on an interval of 30 seconds. Each student observed was observed for 30 minutes. On-task behavior was computed as percentage of times sampled on-task during actual test performance and on-task behavior while directions were being given. On-task behavior during directions was defined as orientation of student's eyes toward either teacher or test booklet and pencil-and-paper compliance with accompanying sample activities. On-task during testing was defined as student's eyes directed toward test booklet, pencil in hand, activity marking, reading, or asking teacher direct questions with specific reference to the test. After completion of the third and final day of testing, students were given an attitude toward tests questionnaire (see Figure 1). This questionnaire consisted of 10 items in an agree/disagree format. Students completed the questionnaire together while the teacher read items to the class.

Insert Figure 1 about here

Results

Achievement

Mean scores on the reading subtest of the CTBS were computed and compared statistically by means of t tests. As can be seen in Table 1, none of the group differences are statistically significant. Interpretation is not possible, however, due to the presence of overwhelming ceiling effects exhibited on all subtests.

Insert Table 1 about here

A supplementary analysis was conducted on the lower half of each group chosen by the previous year's total reading scores and is given in Table 2. This analysis indicates that standardized gain scores between second and third grade testing were significantly higher in favor of the treatment group on Word Attack Subtest and Total Reading Score.

Insert Table 2 about here

On-Task Behavior

Mean on-task behavior during directions, during testing, and total is given in Table 1. As can be seen, no significant group differences were found.

Attitudes Toward Tests

Reliability of the attitude measure was computed by means of a Kuder-Richardson 20 formula and was given at .88, indicating a moderately strong degree of internal consistency for a measure of this type. Differences between the mean scores of the two groups were nonsignificant, _t_ less than 1 in absolute value. An inspection of Figure 2, however, shows that the distribution of these two groups differs strongly. These differences are most

Insert Figure 2 about here

obvious when one employs a curve-smoothing technique of combining the mean scores for each of two adjacent frequencies and are given in the same figure. The difference between these dispersions was tested statistically in two ways: mean differences from the mean in standard scores were computed for subjects in each group and compared statistically. The mean distance from the mean of the placebo group was statistically greater than the average distance from the mean in the training group ($\underline{p} < .01$). In addition, a Kolmogorov-Smirnov two-sample test (Siegel, 1956) was applied to each half of the distribution. For the lower half of each distribution (that is, students scoring 0 through 5 on the measure), the distributions were statistically different ($\underline{Z} = 1.529$, $\underline{p} < .02$), while the upper half of each distribution was not seen to differ significantly ($\underline{Z} = .756$, $\underline{p} = .617$).

Discussion

The present investigation does not offer conclusive evidence that the particular training package employed significantly improved test scores, due to the ceiling effects reported in the Results section. However, it was found that students in the lower half of the treatment group exhibited statistically higher gain scores over the previous year's testing than did the lower half of the placebo group. Particularly, this type of training has previously been seen to demonstrate a significant effect on a subtest similar to the Word Attack subtest in a sample of learning disabled and behaviorally disordered children (Scruggs & Mastropieri, in press.)

That achievement test coaching results in greater levels of on-task behavior on the part of students was not supported by the present investigation. Student on-task behaviors while listening to directions and while taking the test itself were very similar.

Analysis of the attitude data did suggest that students in the treatment group reported more "normal" attitudes than those in the placebo group. The abnormal distribution of scores in the placebo group is highly reminiscent of that of a population under stress (see Wilson, 1973). The fact that the abnormally high number of very negative attitudes was not present in the treatment condition while the number of strongly positive attitudes was relatively similar suggests that this treatment may have contributed to more positive attitudes on the part of those students who may otherwise have developed strong negative reactions to the test and the test-taking situation. It should be noted here that completely positive attitudes toward tests was not expected and is not necessarily a realistic expectation. What was expected was a roughly normal distribution centering around the mean of about 5, which is in fact the distribution seen in the training group. The large proportion of extreme scores in the placebo group (with fully two-thirds of the scores within 1 point of 0 or 10) suggests that the population had been subjected to some stress and had reported widely polarized views on the test-taking process. In the training group, these attitudes seemed to have been ameliorated substantially.

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. Review of Educational Research, 53, 571-585.

Fueyo, V. (1977). Training test-taking skills: A critical analysis. Psychology in the Schools, 14, 180-184.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49, 252-279.

Siegel, S. (1956). Non-parametric statistics. New York: McGraw-Hill.

Scruggs, T. E., Bennion, K., & White, K. R. (1984). Improving achievement test scores in the elementary grades by coaching: A meta-analysis. Unpublished manuscript, Utah State University.

Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled children. Exceptional Children.

Williams, J. (1984). Super score: Training materials for the CTBS. Unpublished materials, Utah State University.

Wilson, G. D. (1973). The concept of conservatism. In G. D. Wilson (Ed.), The psychology of conservatism. New York: Academic Press.

Footnote

Preparation of this manuscript was supported in part by Department of Education Grant #G008300008, Research in the Education of the Handicapped. The authors would like to thank Clyde Bartlett, Principal, and Loila Anderson and Edna Eams, teachers, at Wilson Elementary School, Logan, Utah. We would also like to thank Marilyn Tinnakul and Mary Ellen Heiner for their assistance in the preparation of this manuscript. Address requests for reprints to: Thomas E. Scruggs, Ph.D., UMC 68, Utah State University, Logan, Utah 84322.

Table 1

T-Tests by Group

CTBS Reading Subtests

Variable	N	$\bar{X}$	SD	T	2-tail prob.
Word attack					
Tx	29	29.79	4.87		
				.05	.959
Cx	29	29.72	5.37		
Vocabulary					
Tx	29	26.31	4.58		
				-.49	.624
Cx	29	26.90	4.47		
Comprehension					
Tx	29	26.48	4.06		
				.79	.434
Cx	29	25.51	5.21		
Total reading					
Tx	29	82.59	12.35		
				.13	.898
Cx	29	82.14	14.04		

Table 1 (continued)

Variable	*N*	$\bar{X}$	*SD*	*T*	2-tail prob.
CTBS total battery					
Tx	29	150.17	24.68	-.60	.549
Cx	29	154.03	24.10		
Attitude toward test-taking					
Tx	29	5.59	2.97	.59	.557
Cx	27	5.04	3.95		
On-task during directions					
Tx	18	45.28	15.78	-.75	.458
Cx	18	50.06	21.89		
On-task during testing					
Tx	18	77.67	16.18	.07	.941
Cx	18	77.28	14.98		
Total on-task					
Tx	18	65.78	14.76	-.45	.656
Cx	18	67.78	11.82		

Table 2

Gain Score Differences Between the Lower Half of Each Group (Chosen by Last Year's Total Reading)

Variable	N	$\bar{X}$	SD	Error	T	Prob.
Word attack						
Tx	12	25.83	39.55	11.42	2.41	.012
Cx	14	-20.86	47.06	12.58		
Vocabulary						
Tx	12	18.67	50.77	14.66	.49	.625
Cx	14	7.93	58.69	15.69		
Comprehension						
Tx	12	53.17	37.96	10.96	1.46	.158
Cx	14	24.79	57.54	15.38		
Total of all subtests						
Tx	12	97.67	52.64	15.20	2.51	.019
Cx	14	11.86	107.92	28.84		

Figure Captions

Figure 1. Attitude measure.

Figure 2. Distribution of attitude scores.

Circle YES or NO.

YES NO 1. Taking a test is my favorite thing to do at school.

YES NO 2. Sometimes I am nervous when I take a test.

YES NO 3. I look forward to taking a test.

YES NO 4. I dislike taking a test when I don't know the answers.

YES NO 5. I wish we had fewer tests.

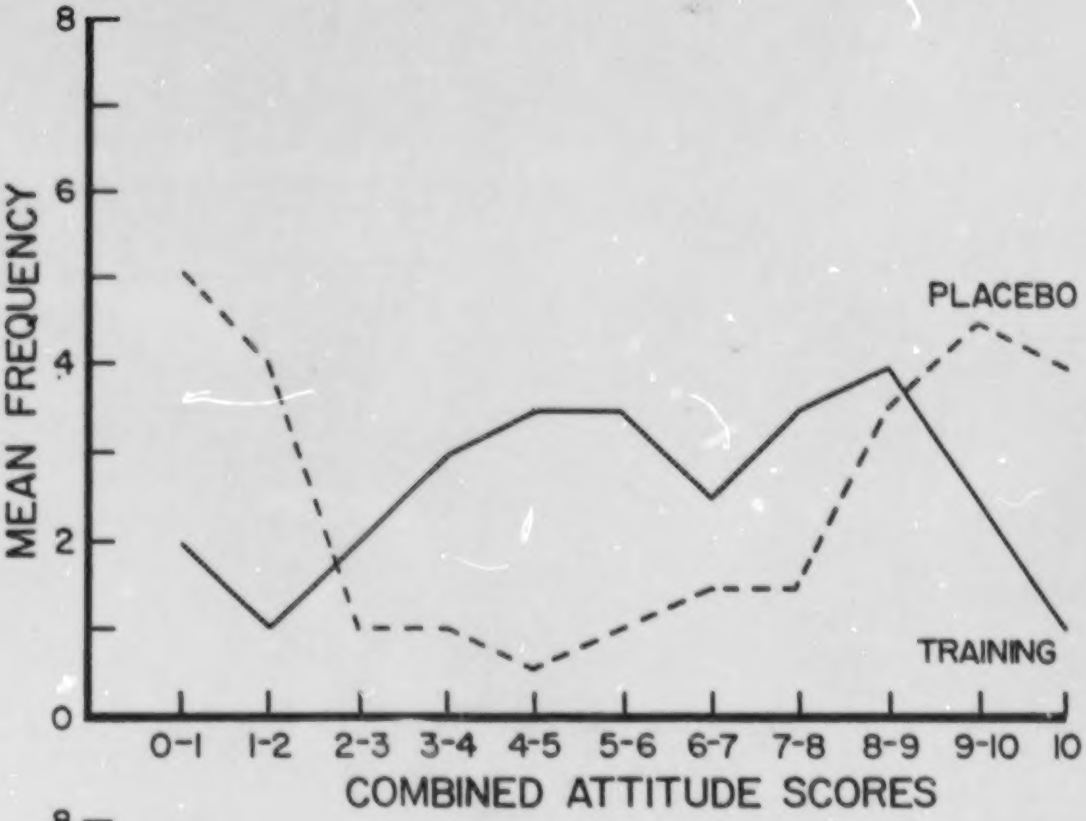
YES NO 6. Taking a test is always fun.

YES NO 7. I like tests even when I don't know the answers.

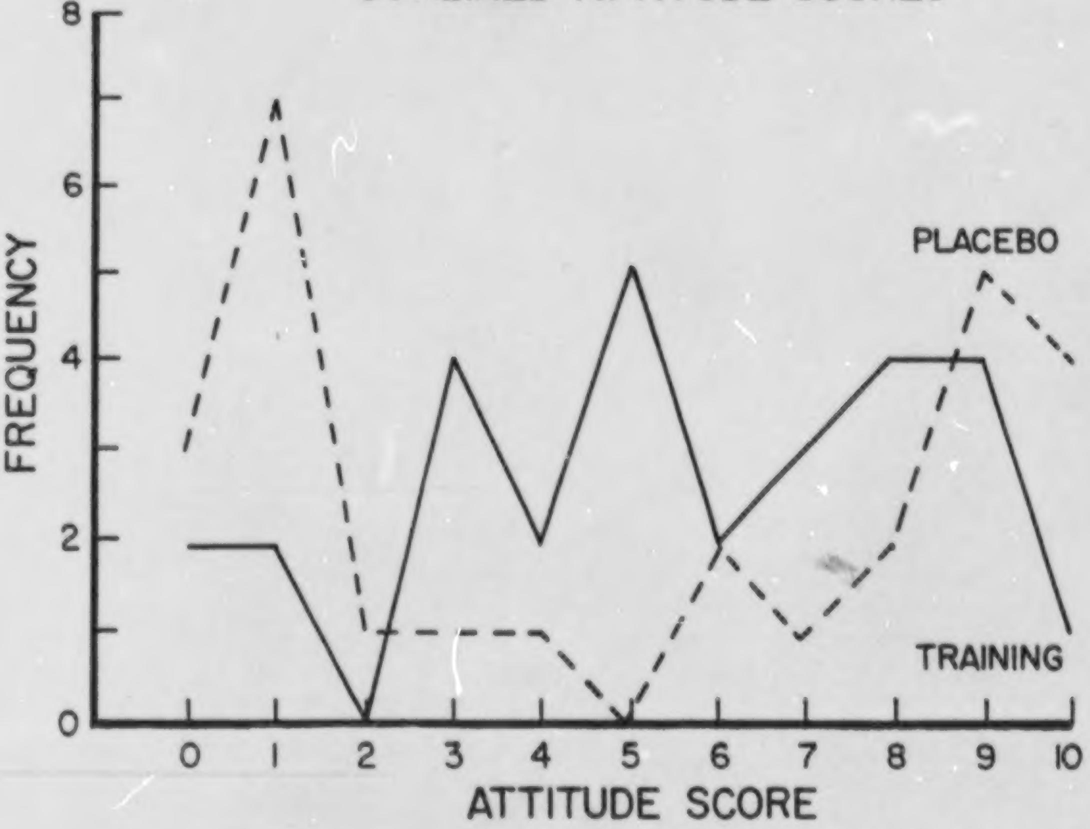
YES NO 8. Taking a test is one of the worst things about school.

YES NO 9. I would rather do something else besides take a test.

YES NO 10. I wish we had more tests.



MEAN FREQUENCY
8
6
4
2
0
PLACEBO
TRAINING
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
10
COMBINED ATTITUDE SCORES



FREQUENCY
8
6
4
2
0
PLACEBO
TRAINING
0
1
2
3
4
5
6
7
8
9
10
ATTITUDE SCORE

APPENDIX L

Improving the Test-Taking Skills of Behaviorally Disordered and Learning Disabled Children

Thomas E. Scruggs
and
Margo A. Mastropieri
Utah State University

Running head: TRAINING TEST-TAKING SKILLS

31;TOM/MANUS/5;IMPROV/TEST

Abstract

Seventy-six third and fourth grade children classified as learning disabled (LD) or behaviorally disordered (BD) were randomly assigned to treatment and control groups. Students assigned to the treatment condition were taught test-taking skills pertinent to reading achievement tests. Students were taught in small groups over a two-week period in such strategies as attending to appropriate stimuli, marking answers carefully, time using, and error avoidance. Following the training procedures, students were administered standardized achievement tests in their normal classroom assignments. Results indicated that trained students scored significantly higher on the Word Study Skills subtest of the Stanford Achievement Test. Scores on the Reading Comprehension subtest were not affected by training. The relevance of these findings to assessment in special education is discussed.

Improving the Test-Taking Skills of Behaviorally Disordered and Learning Disabled Children

Successful performance in school is to a great extent dependent upon the application of effective learning and problem-solving strategies on academic tasks. Students are often called upon to meet particular format and task demands of academic assignments with effective strategies for dealing with these tasks and successfully completing them. Much of the failure of learning disabled (LD) students in school-related tasks has been attributed to a lack of ability in applying such problem-solving strategies (Reid & Hresko, 1980). A body of literature has been established in recent years which documents difficulties of learning disabled students in employing appropriate learning and problem-solving strategies in school. Particular deficits have been noted in the areas of: (a) attending to the critical components of a task (Atkinson & Seunath, 1973; Hallahan & Reeve, 1980; Hallahan, Kauffman, & Ball, 1973; Ross, 1976; Tarver, Hallahan, Kauffman, & Ball, 1976), (b) selecting a strategy appropriate to addressing a particular academic task (Mastropieri, Scruggs, & Levin, in press; Torgesen, 1977; Torgesen & Goldman, 1977), and (c) effectively employing appropriate problem-solving strategies (Hallahan, 1975; Spring & Capps, 1974; Torgeson, Murphy, & Ivey, 1979).

Given the above documented deficiencies, it would appear that one area of particular difficulty for learning disabled and perhaps other mildly handicapped children would be the attentional and problem-solving strategies necessary for successful completion of standardized achievement tests. In these group-administered tests, learners are typically expected to function individually in large-group situations, effectively employ time constraints, and develop and employ strategies specifically suited to answering questions which may be ambiguous or to which the answers are often not completely known (Haney & Scott, 1980). Some recent research with learning disabled students indicates that these students do, in fact, exhibit deficiencies with respect to use of effective strategies in standardized test-taking situations. Scruggs and Lifson (1985) administered questions from standardized reading comprehension tests to LD and non-LD students without providing the accompanying reading passages. Their results indicated that, although non-LD students were able to answer most "reading comprehension" questions without reading the accompanying passages, LD students were less successful. This investigation reiterated previously asked questions concerning what reading comprehension tests actually measure, and also suggested that many LD students may lack some specific test-taking strategies, such as effective use of partial and/or prior knowledge, error avoidance, and elimination strategies. Drawing upon a previous investigation with mostly nondisabled children (Scruggs, Bennion, & Lifson, in press a), Scruggs, Bennion, and Lifson (in press b) recently interviewed learning disabled and non-disabled children with respect to the manner in which they had interpreted and answered reading achievement test items. Analysis of these strategy reports indicated that (a) LD students were less likely to select and utilize strategies appropriate to different types of test questions, and (b) LD students were more likely to be negatively influenced by misleading distractors. Such results suggested that learning disabled and perhaps other mildly handicapped populations may have more difficulty than other students adapting to specific task and format demands of standardized achievement tests and, consequently, resulting scores may be less valid estimations of potential performance than those of other students. Although any observed deficit in "test-taking strategies" on the part of mildly handicapped children would be expected to be representative of more global problem-solving strategy deficits in school-related tasks on the whole, it may be possible that specific training in test-taking skills may be particularly beneficial to children referred for learning and/or behavior problems. Scruggs, Bennion, & Lifson (in press b) hypothesized that, due to differences in format and strategy demands, strategies appropriate for word analysis subtests may be more easily trained than strategies appropriate for reading comprehension subtests.

Previous attempts have been made to improve achievement test scores in regular classrooms by coaching in test-taking skills, but the results have been somewhat mixed and seem to have had a differential effect on different populations. Scruggs, Bennion, and White (in press), in a recent meta-analysis, reported that students from the primary grade levels and students from low socioeconomic backgrounds tended to differentially benefit from extended training in test-taking skills. This finding does suggest that mildly handicapped students may also benefit from instruction in some of the critical skills they apparently lack when confronted with standardized achievement tests.

Scruggs (1984) recently reported the training of test-taking skills to a small sample of LD children. After eight training sessions had been completed, experimental and control students were administered a reduced version of the Stanford Achievement Test (SAT), reading subtests. Results indicated that the experimental students gained significantly on a pre-post criterion measure of test-taking skills, and scored significantly higher (according to a non-parametric test of ranks) on the shortened SAT subtests. Although these results are encouraging, several questions remain. First, could a larger group of mildly handicapped children, including behaviorally disordered (BD) students, be shown to gain from such training? Second, would this training transfer to a standardized administration of the SAT? Finally, if this training could be shown to be successful, it would be interesting to know the actual size of the effect in percentile points, so that an estimate of the practical importance of the treatment could be made. It was the purpose of the present investigation to address these issues.

Method

Subjects

Subjects were 76 third and fourth grade students attending resource rooms or self-contained classes in a large western metropolitan school district.[1] Forty students were third graders and 36 were attending fourth grade classes; 54 of the subjects were boys and 22 were girls. Reading achievement test data are given in the "Results" section. Fifty students were classified as BD, and 26 students were classified as LD according to federal, state, and local school district criteria. For behavioral disorders, the definition included students whose behavioral or emotional functioning over time adversly affected educational performance and required special education service. For learning disabilities, the definition included a 40% discrepancy between ability and achievement. Although specific academic deficiencies were not criteria for BD classification, a separate analysis of achievement scores of LD and BD children in this particular district indicated that differences in academic achievement between the two groups were trivial (Scruggs & Mastropieri, 1984). Eighteen students were enrolled in self-contained classes, and 58 students were attending resource rooms. Subjects were stratified by grade level and randomly assigned to experimental and control groups, without regard to category of exceptionality.

Materials

Materials were developed as part of a larger project involving improving test-taking skills of LD and BD elementary students (Taylor & Scruggs, 1983) and consisted of eight scripted lessons for each grade level in a direct instruction format and accompanying workbooks for students which included pencil-and-paper practice activities (exact materials used are given in Scruggs & Williams, in press). The general test-taking strategies taught in these materials included attending to directions, marking answers carefully, choosing the best answer carefully, error avoidance strategies, and appropriate situations for soliciting teacher attention. In addition, specific test-taking strategies were taught for each reading subtest in the Stanford Achievement Test. These included structured practice in specific test formats for each subtest and specific application of general test-taking strategies to each specific subtest. For example, with respect to the letter-sound subtest, students were taught to employ the following sequence of strategies:

1. Read the first word.
2. Pronounce to yourself and think of the sound of the underlined letter.

3. Carefully look at all the answer choices and choose the word with the same sound as the underlined letter.

4. If you don't know all the words, read the words you do know, or read parts of individual words that you may know.

5. If you are not sure of the answer, see if there are some answers that you are sure are not correct, and eliminate those.

6. Color in the answer quick, dark, and inside the line.

7. Guess if you are not sure; never skip an answer.

Procedure

Experimental subjects were taught by four trained experimenters in small groups ranging from one to five in size. Four 20-30-minute lessons were given per week for two weeks. Positive responding and attention to task were reinforced with stickers. Immediately prior to the training sessions, and immediately after the last training session, students were administered a criterion test of the skills which were taught. This test was a 10-item test of test-taking skills including questions about time using, question asking, and elimination strategies. The first seven sessions taught the use of test-taking strategies within the specific context of each of the reading-related subtests. The last session consisted of a general review of all previous procedures. Each day of instruction involved extensive work with practice activities applied to practice test items. At no time during this training procedure were subjects taught any information concerning the content of the test which was not given in the published test directions. Within five days of completion of the training sessions, students were administered the Stanford Achievement Test. This administration was done in the regular or self-contained classroom settings by their regularly assigned teachers. Although teachers were aware of the membership of each student in the experimental group, response protocols were scored by machine. Results

Pre and posttests of the experimental students on the criterion measure were compared statistically by means of a correlated t test. It was found that the performance on the posttest was significantly higher than pretest scores ($p < .01$). Students scored an average of 40% percent correct on the pretest, and 77% correct on the posttest.

Eight students (5 experimental and 3 control) did not complete either or both subtest of the SAT and were excluded from further analysis. Experimental students scored an average of the 25.3 percentile (SD = 20.0) on the Word Study Skills Subtest and the 16.8 percentile (SD = 15.0) on the Reading Comprehension Subtest of the SAT. Control subjects scored an average of the 17.4 percentile on Word Study Skills (SD = 18.3) and the 16.4 percentile (SD = 15.0) on Reading Comprehension. Student percentile scores were entered into a 2 (group) X 2 (subtest) analysis of variance (ANOVA), with repeated measures on the subtest variable (Winer, 1971), which yielded significance on subtests, F(1,66) = 4.96, p<.03, and group X subtest interaction, F(1,66) = 7.06, p < .01. The main overall effect by group was not statistically significant, F(1,66) = 1.21, p < .30. Analysis of simple effects (Winer, 1971) indicated that experimental and control students differed significantly with respect to the Word Study Skills subtest, _t_(66) = 2.07, p < .05, but not the Reading Comprehension subtest, _t_(66) = -.15, p > .20. The group X subtest interaction is depicted graphically in Figure 1.

Insert Figure 1 about here

Discussion

The analysis of pre and posttest scores indicated that test-taking skills could be successfully taught to this sample of third and fourth grade mildly handicapped children. The fact that significant gains were made in these critical skills suggests that mildly handicapped children at this age level do lack certain test-taking skills which are potentially useful in taking standardized achievement tests.

Analysis of the test data indicated that training in test-taking skills did significantly increase scores on the Word Study Skills Subtest of the Stanford Achievement Test for this sample of mildly handicapped students. The overall effect size for this investigation, .20, is twice as large as the mean effect size found for similar investigations with elementary school aged non-handicapped children (Scruggs, Bennion, and White, in press), but similar to that obtained for primary grade students under conditions of extended training (for this age group, an effect size of .10 is equivalent to approximately one month of academic achievement). The effect size of .43 for the Word Study Skills subtest is comparable to the mean effect size found for children of low socioeconomic status (SES) under conditions of extended training, but much higher than mean effect sizes found for higher SES children, or lower SES children with shorter training periods (Scruggs, Bennion, & White, in press).

As predicted by recent research (Scruggs, Bennion, & Lifson, in press b), performance was increased on the Word Study Skills subtest and not the Reading Comprehension subtest. The fact that the Word Study Skills subtest was increased significantly may be a function of the fact that this particular subtest involves many format changes over a short period of time, and thus was more amenable to increased performance through guided practice and feedback on successful skills necessary for completion of the subtest. Strategy deficits previously observed on the Reading Comprehension subtest, however, were not thought to be easily remediable. These deficits included ineffective use of deductive reasoning strategies, inability to distinguish between recall and inferential questions, and inappropriate levels of confidence in answer choices (Scruggs, Bennion, & Lifson, in press b).

The finding of positive training effects replicates that of Scruggs (1984), and extends it to a larger population representing different categories of exceptionality on a standardized test administration. Although the present results are encouraging, several questions remain. First, students in this investigation were trained by project personnel in order to insure fidelity of treatment. The extent to which teacher implementation would effect results is not known.[2] Second, the overall sample size, the fact that subjects were not stratified by category of exceptionally, and the disproportionally small number of LD students in the present sample did not allow sufficient power (Cohen, 1969) to separately assess the effects for LD vs BD students, although it may be interesting to do so in future research. Also, it is not certain which training procedures were most responsible for the observed effects. It is likely, however, that training in strategies needed for meeting specific format demands was more beneficial than the training given in general test-taking strategies (e.g., time-using strategies), for the reason that a different effect was observed on the two subtests. Finally, the extent to which such training can benefit different grade levels and content areas (such as math) remain to be seen.

The present authors are currently investigating such possibilities (Taylor & Scruggs, 1983).

The usefulness of standardized achievement tests in special education has been, and remains, a controversial issue (see Salvia & Ysseldyke, 1981) not intended to be addressed by the results of of the present investigation. It must be considered, however, that the observed effect (that of raising mean scores from the 17th to the 25th percentile) could be sufficient to prevent special education referral for some students in schools where such test scores are weighted heavily. The present authors do not subscribe to the notion that special educational services are undesirable, and that students should be "saved" from them whenever possible. It is our view that referral for special education services is a serious procedure which must take into account many different considerations, both qualitative and quantitative, and for which the ultimate goal must be optimal educational service delivery for the individual child. If standardized achievement tests are to be used for this purpose, then it is important that the score obtained be as nearly as possible a reflection of the child's knowledge of the content area being assessed[3]. To this end, training in test-taking skills may be useful. There are other ends, however, which we feel ought to be considered in such training. Since the skills trained in the present investigation apparently did transfer to a standardized test situation, it seems likely that similar training may generalize to other related tasks, e.g., for older students, taking a driver's test or an aptitute test relevant to a specific employment opportunity.

Finally, test taking can be viewed simply as a common task in todays' schools, but not a particularly pleasant experience to a mildly handicapped student who typically performs poorly, or who does not fully understand testing conventions and formats. In this case, training in test-taking skills could be regarded as another means to improve the ability of the individual child to function in the outside world, a goal to which all special educators aspire.

References

Atkinson, B. R., & Seunath, O. H. M. (1973). The effect of stimulus change in attending behavior in normal children and children with learning disorders. Journal of Learning Disabilities, 6, 569-573.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Hallahan, D. P. (1975). Comparative research studies on the psychological characteristics of learning disabled children. In W. M. Cruickshank & D. P. Hallahan (Eds.), Perceptual and learning disabilities in children, Vol. 1. Psychoeducational practices. Syracuse, NY: Syracuse University Press.

Hallahan, D. P., & Reeve, R. E. (1980). Selective attention and distractibility. In B. Keogh (Ed.), Advances in special education (Vol. 1). Greenwich, CT: Jai Press.

Hallahan, D. P., Kauffman, J. M., & Ball, D. W. (1973). Selective attention and cognitive tempo of low achieving and high achieving sixth grade males. Perceptual and Motor Skills, 36, 579-583.

Haney, W., & Scott, L. (1980). Talking with children about tests: A pilot study of test item ambiguity. National Consortium of Testing Staff Circular No. 7. Cambridge, MA: The Huron Institute.

Mastropieri, M. A., Scruggs, T. E., & Levin, J. R. (in press). Memory strategy instruction with learning disabled adolescents. _Journal of Learning Disabilities_.

Reid, D. K., & Hresko, W. P. (1980). Thinking about thinking about it in that way: Test data and instruction. _Exceptional Education Quarterly_, _1_(3), 47-57.

Ross, A. O. (1976). _Psychological aspects of learning disabilities and reading disorders_. New York: McGraw-Hill.

Salva, J., & Ysseldyke, J. E. (1981). _Assessment in remedial and special education_ (2nd ed.). Boston: Houghton Mifflin.

Scruggs, T. E. (1984). _Improving the test-taking skills of learning disabled students_. Unpublished manuscript, Utah State University, Logan.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press a). An analysis of children's strategy use on reading achievement tests. _Elementary School Journal_.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press b). Spontaneously employed test-taking strategies of learning disabled students on reading achievement tests. _Learning Disability Quarterly_.

Scruggs, T. E., Bennion, K., & White, K. (in press). Improving achievement test scores in the elementary grades by coaching: A meta-analysis. In Scruggs, T. E., *Administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementry school children*. Final report. (ERIC Document Reproduction Service)

Scruggs, T. E., & Lifson, S. (1985, April). *Are learning disabled students "test-wise?": An inquiry into reading comprehension test items*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Scruggs, T. E., & Mastropieri, M. A. (1984, November). *Academic characteristics of behaviorally disordered students*. Paper presented at the Conference on Severe Behavior Disorders of Children and Youth, Tempe, AZ.

Scruggs, T. E., & Williams, J. (in press). *Teaching test-taking skills to learning disabled and behaviorally disordered children. Super score: Test taking manuals and workbooks*. (ERIC Document Reproduction service)

Spring, C., & Capps, C. (1974). Encoding speed, rehearsal, and probed recall of dyslexic boys. *Journal of Educational Psychology*, *66*, 780-786.

Tarver, S. G., Hallahan, D. P., Kauffman, J. M., & Ball, D. W. (1976). Verbal rehearsal and selective attention in children with learning disabilities: A developmental lag. *Journal of Experimental Child Psychology*, *22*, 375-385.

Taylor, C., & Scruggs, T. E. (1983). Research in progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 50, 277.

Torgesen, J. K. (1977). The role of nonspecific factors in the task performance of learning disabled children: A theoretical assessment. Journal of Learning Disabilities, 10, 27-34.

Torgesen, J. K., & Goldman, T. (1977). Verbal rehearsal and short-term memory in reading-disabled children. Child Development, 48, 56-60.

Torgesen, J. K., Murphy, H. A., & Ivey, C. (1979). The influence of an orienting task on the memory performance of children with reading problems. Journal of Learning Disabilities, 12, 396-401.

Winer, B. J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

Ysseldyke, J. E., Algozzine, B., Richey, L., & Graden, J. (1982). Declaring students eligible for learning disability services: Why bother with the data? Learning Disability Quarterly, 5, 77-44.

Footnote

The preparation of this manuscript was supported in part by a grant from the Department of Education, Special Education Programs, #G008300008. The authors would like to thank Dr. Joyce Barnes and the teachers and administrators of the Granite School District for their cooperation and assistance. The authors would also like to thank Marilyn Tinnakul and Mary Ellen Heiner for their assistance in the preparation of this manuscript.

[1] A group of second grade LD and BD students was initially intended for inclusion in this study, but was dropped due to methodological problems involving sample selection and subject attrition.

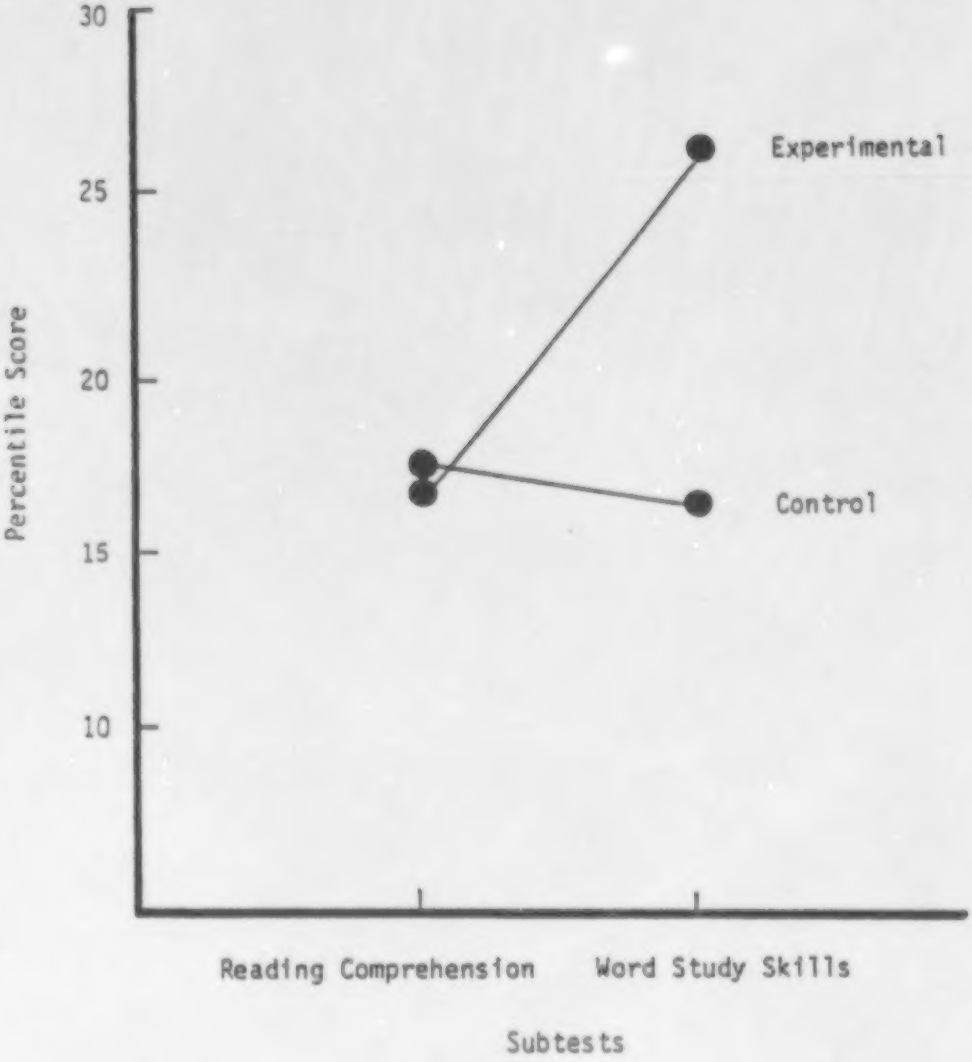
[2] An argument can be made that, since control subjects did not receive a 'placebo' treatment (i.e., non-instructional contact with the experimenters for an equivalent trial period), the observed effects may be due to a reaction to the novelty of experimenter contact and not the training procedure. A decision was made not to deliver placebo training to the control group so that control subjects would have received additional teacher-led instruction as the comparison treatment, and so that their instructional time would not have been wasted on non-educational treatments. Furthermore, the "novelty" argument seems untenable because: (a) a recent meta-analysis by the present authors indicated that such subtle treatments were highly unlikely to

raise test scores, and (b) such an argument does not explain why only one, and not both, subtest scores were raised.

[3] In fact, a question has been raised concerning to what extent any assessment data are used for making placement decisions. (see Ysseldyke, Algozzine, Richey, & Graden, 1982, for a discussion of this issue).

Figure Caption

Figure 1. Group by subject interaction.


30
25
20
15
10
Percentile Score
Experimental
Control
Reading Comprehension
Word Study Skills
Subtests

APPENDIX M

Can LD Students Effectively Use Separate Answer Sheets?

Debra Tolfa and Thomas E. Scruggs

Utah State University

Running head: SEPARATE ANSWER SHEETS

DISK 26; TOM/MANUS; SEPARATE/ANS

Abstract

One hundred three regular class and learning disabled (LD) students were administered three subtests of the Comprehensive Test of Basic Skills for which all correct answers had been identified in the student test booklet. Analysis of the completed separate answer sheets indicated that LD students answered fewer total items than their non-disabled counterparts, but did not differ with respect to percent of items answered correctly. In addition, descriptive but non-significant differences were found for number of answer spaces filled in outside the line. Implications for training and assessment are given.

Can LD Students Effectively Use Separate Answer Sheets?

Introduction

In recent years, research attention has focused upon the skills and strategies learning disabled (LD) students apply indepenently to test-taking situations (Taylor & Scruggs, 1983). Any observed deficiencies in these "test-taking skills" could be considered (a) a potential source of measurement error (e.g., Ebel, 1965), as well as (b) a potential area for needed intervention. And, although research has indicated that group administered achievement tests are reliable and valid for LD students (e.g., Price, 1984), some deficiencies in test-taking skills have been observed in this population. Scruggs and Lifson (1985) administered reading comprehension questions to LD and nondisabled students without providing the accompanying reading passages. They found that although nondisabled readers were apparently able to make use of such strategies as partial and/or prior knowledge, error avoidance, elimination, and use of information from other test items, LD students were much less successful. Drawing upon a previous investigation with mostly nondisabled students (Scruggs, Bennion, & Lifson, 1985), Scruggs, Bennion, and Lifson (in press) recently interviewed LD and nondisabled students concerning the "test-taking strategies" they spontaneously employed on reading achievement tests. It was concluded that (a) LD students were less successful at selecting strategies appropriate for different types of test questions, and (b) LD students were less successful at adapting to novel test formats. Given the number and frequency of format changes on standardized achievement tests, these factors could exert a potentially strong influence on LD students' test performance (Tolfa, Scruggs, & Bennion, in press).

Another important format change which takes place on standardized tests after the primary grades, is the inclusion of separate answer sheets to facilitate machine scoring. The ability to use separate answer sheets appears to be developmental in nature, with students in grades one and two showing better performance levels when test booklets are used as compared with separate answer sheets (Ramseyer & Cashen, 1971). Cashen and Ramseyer (1969) indicated that the need for use of the test booklet marking decreases as the grade level of the student increases. Typically, standardized tests begin the use of separate answer sheets in grade four. The implications for the fourth or fifth grade learning disabled student functioning two years behind his peers in perceptual-motor skills become obvious.

It has been suggested that students can be trained in the skill of separate answer sheet usage (McKee, 1967; Ramseyer & Cashen, 1971). McKee (1967) described training third graders to successfully use separate answer sheets. However, this study represented more of a subjective evaluation than a tightly designed research study. Ramseyer and Cashen (1971) concluded that first and second graders were unable to utilize separate answer sheets effectively even after practice sessions. Both studies (McKee, 1967; Ramseyer & Cashen, 1971) were conducted with students functioning in regular classrooms.

The present investigation examined the effects of separate answer sheet usage with fourth grade learning disabled students. The study was conducted to determine if, in fact, fourth grade learning disabled students were functioning less efficiently than their normally functioning peers in the use of the separate answer sheet, with relative ability to answer test items controlled.

Method

Subjects

Subjects were 103 fourth grade students enrolled in elementary schools in a rural university community in northern Utah. All students were enrolled in the fourth grade. Nineteen of these students (14 boys and 5 girls) were classified as learning disabled according to P.L. 94-142 and Utah State guidelines, which include average ability coupled with two years discrepancy on standardized achievement tests. Average Wecshler Intelligence Scale for Children-Revised (WISC-R) for the LD group was 97.94 (SD = 8.81); Average Total Reading grade equivalent score from the Woodcock-Johnson was 2.63 (SD = .90) for the LD students. Eighty-four (48 boys and 36 girls) nondisabled students were functioning within the regular classroom setting. These students were functioning at or near grade level, and had not been identified as "gifted," "remedial," or identified for special services of any kind. Average Total Reading grade equivalent from the California Test of Basic Skills was 4.24 (SD = 1.42).

Materials

Experimental materials consisted of the test booklet appropriate for the fourth grade Comprehensive Test of Basic Skills (CTBS) and the CTBS fourth grade answer sheet. All correct responses had been marked with a black arrow in the test booklet. Subtests one, five, and seven were selected as target subtests. All subtests contained 45 questions. A presenter's script was prepared.

Procedure

Nineteen learning disabled students and 84 regular class fourth graders were administered the three subtests by one of three examiners. Examiners were given a written script to ensure all students received the same directions. All students were administered the assignment in a group setting with the exception of three LD students who were administered the exercise individually in their resource room setting.

Students were told that they would be given a test that already had the correct answers marked and that their task was to mark the correct answers on the separate answer sheet. They were told to work as quickly and carefully as possible; they would be given three minutes to work on each subtest. Students and examiners worked the examples together, and examiners checked to ensure students were completing the correct subtest sections on the answer sheet.

Answer sheets were scored by recording number of items completed, number of items answered correctly, and number of items marked outside the established 5 mm radius from the center of each answer circle for each subtest. This distance represented the point at which the pencil mark could intrude into an adjacent answer space.

Results

Each subtest was evaluated based on total number of items completed, total percent marked correctly, and total percent marked outside the circle (i.e., more than 5 mm from the center). For total completed, students in the nondisabled group obtained a mean score of 96.65 (SD = 18.8), while students in the learning disabled group obtained a mean score of 86.2 (SD = 18.0). These differences were statistically significant in favor of the nondisabled group, $\underline{t}(99) = 2.19$, $p = .03$. For percent of marked items answered correctly, however, differences were not observed. Students in the nondisabled group recorded 98% (SD = .06) of their answers correctly, while LD students marked 96% (SD = .13) of their answers correctly. Because obtained variance differed for the two groups ($p < .01$), a separate variance estimate was used, with a correction for degrees of freedom (Ferguson, 1982) which yielded a $t(20) = .61$, $p = .55$.

In addition, a descriptive, non-significant difference was found when groups were compared with respect to percent of answer spaces marked outside the line, $t(21) = 1.71$, $p = .10$ (separate variance estimate). Descriptively, the nondisabled group marked an average percent of 7.8 (SD = 8.6) answers outside the line, while the LD sample marked an average percent of 13.0 (SD = 17.7) answers outside the line, assessed as a function of total number of answers marked.

Discussion

LD students were seen to differ significantly from nondisabled students with respect to ability to utilize a separate answer sheet in answering standardized achievement test questions. These differences were most pronounced in the area of speed and less pronounced in the area of accuracy and neatness, although descriptive differences were also found in these areas. The present data strongly suggested that the achievement test performance of LD students may be differentially hampered in performance by separate answer sheets, resulting in increased measurement error. Further research is needed, however, to document the exact extent performance may be inhibited under standardized test administration conditions.

Two possible interventions can be imagined to help correct such possible difficulties: One possibility is to modify the tests themselves, while the other possibility is to train LD students to be more efficient with separate answer sheets. And, in fact, such procedures have recently received attention in the research literature. Beattie, Grise, and Algozzine (1982) assessed the effectiveness of several test modifications, including imbedding the answer circle within the test booklet, on the competency test performance of LD students. Although some descriptive advantages were noted, the overall modifications failed to produce any strong consistent effect. With respect to the second possibility, attempts to train LD children in use of novel test formats, including separate answer sheets, have been successful. Scruggs and Tolfa (1985) and Scruggs and Mastropieri (in press), reported successfully teaching such 'test-taking skills' to LD students, to the extent that test performance, subsequent to training, was significantly higher than that of untrained controls. The fact that resulting effect sizes in these investigations were higher than those usually reported in the literature (Scruggs, Bennion, & White, in press) supports the notion that LD students may indeed demonstrate relative deficits in a variety of 'test-taking skills' (Scruggs & Lifson, in press). Further research can do much to further describe the nature of such deficits, and develop effective means of remediation. The present authors are, in fact, currently engaged in such an effort (Taylor & Scruggs, 1983).

References

Beattie, S., Grise, P., & Algozzine, B. (1982). Effects of test modifications on minimum competency test performance of third grade learning disabled students. (ERIC Document Reproduction Service No. ED 226 046)

Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall.

Ferguson G. A., (1982). Statistical analysis in psychology and education (5th ed.). New York: McGraw Hill.

Price, P. A. (1984). A comparative study of the California Achievement Test and the Key Math Diagnostic Arithmetic Test with secondary LH students. Journal of Learning Disabilities, 17, 392-396.

Scruggs, T. E., Bennion, K., & Lifson, S. (1985). An analysis of children's strategy use on reading achievement tests. Elementary School Journal, 85, 479-484.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press). Learning disabled students spontaneous use of test-taking skills on reading achievement tests. Learning Disability Quarterly

Scruggs, T. E., Bennion, K., & White, K. R. (in press). Improving achievement test scores in the elementary grades by coaching: A meta-analysis. Elementary School Journal.

Scruggs, T. E., & Lifson, S. A. (1985, April). Are learning disabled students 'test-wise?' An inquiry into reading comprehension test items. Paper presented at the annual meeting of the American Educational Research Association.

Scruggs, T. E., & Lifson, S. (in press). Current conceptions of test-wiseness: Myths and realities. School Psychology Review.

Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled students. Exceptional Children.

Scruggs, T. E., & Tolfa, D. (1985). Improving the test-taking skills of learning disabled students. Perceptual and Motor Skills, 60, 847-850.

Taylor, C., & Scruggs, T. E. (1983). Research in progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 50, 277.

Tolfa, D., Scruggs, T. E., & Bennion, K. (in press). Format changes in reading achievement tests: Implications for learning disabled students. Psychology in the Schools.

Author Notes

The research reported here was supported in part by a grant from the Department of Education, Office of Special Education Programs, #G008300008. The authors would like to thank Mrs. Bonnie Olsen for her assistance with this project and Mary Ellen Heiner for her assistance in the preparation of the manuscript. Address requests for reprints to Thomas E. Scruggs, Developmental Center for Handicapped Persons, UMC 68, Utah State University, Logan, UT 84322.

APPENDIX N

Attitudes of Behaviorally Disordered Students Toward Tests: A Replication

Debra Tolfa, Thomas E. Scruggs, and Margo A. Mastropieri

Utah State University

Running head: TEST ATTITUDES II

DISK 88; TOM/MANUS; TEST/REPLICA

Abstract

Ninety-six behaviorally disordered and more average students were administered a test attitude survey immediately after district-wide standardized achievement testing. Results were consistent with previous research which suggested behaviorally disordered students may report lower attitudes than their more typical peers. In addition, differentially lower scores were found for behaviorally disordered girls, while no sex differences were found in the more average group.

Attitudes of Behaviorally Disordered Students toward Tests: A Replication

The behaviorally disordered student is thus classified based on average or near average intellectual ability in addition to social or emotional functioning that is substantially different from other students the same age. Behaviorally disordered students have repeatedly shown academic deficiencies (Mastropieri, Jenkins, & Scruggs, 1985; Motto & Wilkins, 1968; Stone & Rowley, 1964). Several variables, including attitude toward school subjects (Silberberg & Silberberg, 1971), impulsivity (Letteri, 1979), and responses toward test-taking situations (Forness & Dvorak, 1982; Scruggs & Mastropieri, in press; Scruggs, Mastropieri, Tolfa, & Jenkins, 1985), have been identified as possible contributing factors to academic deficiencies.

The present study investigates the behaviorally disordered student's attitude toward test-taking situations. In the Scruggs et al. (1985) study, conflicting results were found. In Study 1, responses of fifth and sixth grade behaviorally disordered students were compared with those of their normally functioning peers on a 12-item test-attitude survey. Results indicated that the behaviorally disordered students differed significantly from their normally functioning peers on the overall survey as well as the specific factors involving subjective feelings about tests and feelings about the personal importance of tests. Groups did not differ with respect to evaluation of the objective value of tests. The sample in this study was relatively small (N = 37), however, and the survey contained too few items to draw firm conclusions.

In Study 2 of the same investigation, 75 regular classroom students and 25 self-contained behaviorally disordered students were administered a longer test attitude survey. Groups, which were equivalent with respect to number, age, sex, and grade, were then compared. There was no difference between groups on the total survey, or on "personal feeling" items, or on "value of tests" items. Scruggs et al. (1985) proposed several possible explanations for these discrepant findings, including that fact that Study 2 was conducted at the beginning of the school year when students had not had much recent experience with test-taking, while Study 1 was conducted at the end of the previous school year after students had recently experienced testing situations.

The present investigation was conducted to help clarify the conflicting results of the Scruggs et al. (1985) investigation. A larger population, including a greater number of grade levels, was compared on a revised version of the test attitude survey utilized in Study 2 of the Scruggs et al. (1985) investigation. In addition, a larger sample of girls was employed in the present investigation so that an evaluation of possible group by sex interaction effects could be made.

Method

Subjects

Subjects were 96 elementary school children attending a public school in a western metropolitan community. Students were enrolled in grades one through six. Forty-eight of these students were classified as behaviorally disordered, while 48 were more typical students enrolled in regular classrooms in the same school. To be included in the study from the regular classroom, students were selected at random, using a stratified random sampling technique, from a population of 122 students representing the same grade levels. When possible, equal numbers of boys and girls per grade level were selected to match numbers represented in the target population. The breakdown by grade level and sex for each group was as follows: three students (1 boy, 2 girls) were enrolled in first grade, eight students (5 boys, 3 girls) in second grade, four students (all boys) in third grade, eight students (6 boys, 2 girls) in fourth grade, eleven students (behaviorally disordered = 9 boys, 2 girls; regular class 6 boys, 5 girls) in fifth grade, and fourteen students (behaviorally disordered = 11 boys, 3 girls; regular = 9 boys, 5 girls) were enrolled in sixth grade.

Students were identified as behaviorally disordered according to state and P.L. 94-142 guidelines, which included students exhibiting behavior or emotional conduct over time which adversely affected educational performance, and required special education services in self-contained classrooms.

Materials and Procedure

The 17-item Test Attitude Survey was constructed based on results from previous investigations which also examined test-taking attitudes of students (Scruggs, Bennion, & Williams, 1985; Scruggs, Mastropieri, Tolfa, & Jenkins, 1985) and contained such items as "tests are an important part of school," "tests are more important to the teacher than to me," "tests are a waste of time," "I try my best when I take a test," and "I do poorly on tests." Items were intended to reveal students' feelings of the importance of tests to themselves and to parents and teachers, as well as their own feelings toward tests.

The measure was administered immediately subsequent to yearly achievement testing. Administration of the survey was conducted in the students' regular classroom, and items were answered together as the teacher read each item aloud. Students were given 1 point for a positive response (i.e., "yes" to a positive statement, or "no" to a negative statement) and 0 points for a negative response.

Results

The reliability (Kuder-Richardson 20) of the present survey for this sample was .81, which was slightly higher than that of previous coefficients of .76 and .75 (Scruggs et al., 1985). Response data were entered into a 2 (group) x 2 (sex) analysis of variance (ANOVA), and yielded significance for groups, $F(1,92) = 19.73$, $p < .001$. No significant main effect was found for sex, $F(1,92) = 2.46$, $p = .12$. Finally, the interaction of group by sex was seen to closely approach significance, $F(1,92) = 3.59$, $p = .06$. Follow-up *t*-tests indicated that girls in the behaviorally disordered groups reported differentially lower attitudes ($t[46] = 3.56$, $p < .001$), while boys and girls in the more average group did not differ ($t < 1$). Descriptively, the more average group reported more positive attitudes than the behaviorally disordered group with mean scores of 14.56 (SD = 2.03) and 12.15 (SD = 3.69), respectively. Sex by group differences are depicted graphically in Figure 1.

Insert Figure 1 about here

A factor analysis of responses for the total group was calculated using the same procedures as in the Scruggs et al. (1985) investigation. In this analysis, however, meaningful factors consisting of more than two or three items were not identifiable. This finding was inconsistent with that of Scruggs et al. (1985, Study 1), which identified three distinct factors: (a) personal importance of tests, (b) objective worth of tests, and (c) personal feelings about tests.

Discussion

The present investigation replicated the findings of Study 1 in Scruggs et al. (1985), and suggested that behaviorally disordered children do report different, less positive attitudes toward test-taking situations than their more normally functioning peers. This study also expanded previous findings to include grades one through six.

Although sample size and matching procedures more closely paralleled Experiment 2 of the Scruggs et al. (1985) investigation, the findings between that study and the present investigation were in opposition. This may suggest that the time of the school year influenced students' responses. While Study 2 in the Scruggs et al. (1985) investigation was conducted during the beginning of the year when students had not recently undergone testing, the present study was conducted following the yearly administration of the standardized tests. The exposure to the testing situation may have given students a more realistic outlook on their test-taking attitude.

Finally, although the sex by grade interaction was not significant by conventional standards, the effect was sufficiently tangible to warrant further investigation.

These results suggest that behaviorally disordered students do differ from their normally functioning peers on test-taking attitudes. Further research could do much to clarify any possible causal relation between test scores and test attitudes of behaviorally disordered students.

References

Forness, S. R., & Dvorak, R. Effects of test time limits on achievement scores of behaviorally disordered adolescents. Behavioral Disorders, 1982, 7, 207-212.

Letteri, C. A. The relationship between cognitive profiles, levels of academic achievement, and behavior problems. Behavior Disorders Monographs, 1979, 2, 74-84.

Mastropieri, M. A., Jenkins, V., & Scruggs, T. E. Academic and intellectual characteristics of behaviorally disordered children. Unpublished manuscript, Utah State University, 1984.

Motto, J. J., & Wilkins, G. A. Educational achievement of institutionalized emotionally disturbed children. Journal of Educational Research, 1968, 61, 218-221.

Scruggs, T. E., Bennion, K., & Williams, N. J. Effects of training in test-taking skills on test performance, attitudes, and on-task behavior of elementary school children. Unpublished manuscript, Utah State University, 1985.

Scruggs, T. E., Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled children. Exceptional Children.

Scruggs, T. E., Mastropieri, M. A., Tolfa, D., & Jenkins, V. Attitudes of behaviorally disordered students toward tests. Perceptual and Motor Skills, 1985, 60, 467-470.

Silberberg, N. E., & Silberberg, M. C. School achievement and delinquency. Review of Educational Research, 1971, 41, 17-32.

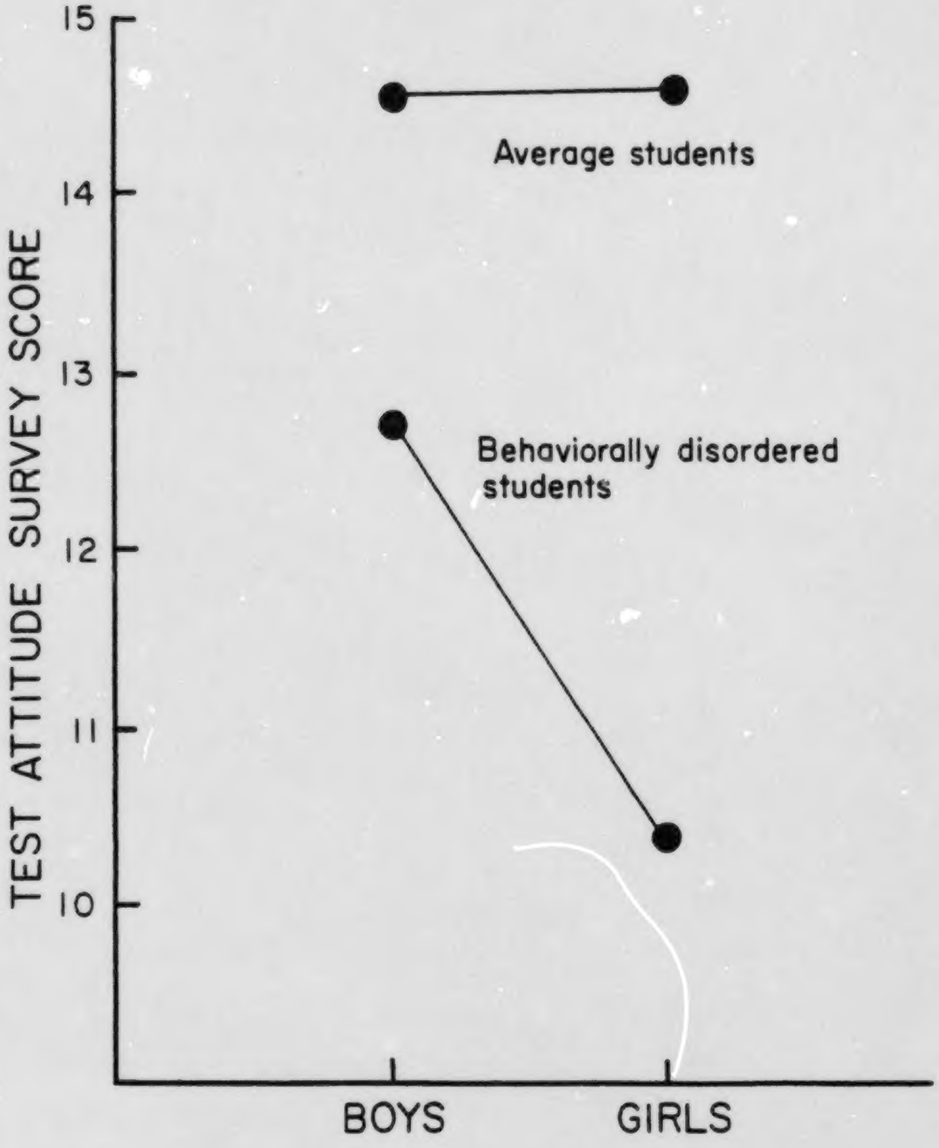
Stone, F. B., & Rowley, V. N. Educational Disability in emotionally disturbed children. Exceptional Children, 1964, 30, 423-426.

Author Notes

The research described here was supported in part by a grant from the Department of Education, Special Education Programs, Office of Special Education and Rehabilitative Services, No. G008300008. The authors would like to thank Ms. Cathy Smith, Coordinator of Special Education, Hillview Elementary School, Salt Lake City, Utah, for her cooperation; as well as Ursula Pimentel and Mary Ellen Heiner for their assistance in the preparation of this manuscript. Address requests for reprints to Thomas E. Scruggs, UMC 68, Utah State University, Logan, UT 84322.

Figure Caption

Figure 1: Sex by group interaction.


15
14
13
12
11
10
TEST ATTITUDE SURVEY SCORE
Average students
Behaviorally disordered students
BOYS
GIRLS

APPENDIX O

The Effects of Coaching on the Standardized Test Performance of Mildly Handicapped Students

Thomas E. Scruggs, Margo A. Mastropieri, Debra Tolfa

Utah State University

Running Head: TEST PERFORMANCE

DISK 55; TOM/MANUS/8; TEST/PERFORM

Abstract

Eighty-five mildly handicapped (learning disabled or behaviorally disordered) students were assigned at random to either a control condition or a condition in which students received five days' training on test-taking skills relevant to the Stanford Achievement Test. Results of test scores indicated that trained students scored significantly higher on tests of reading decoding and math concepts. A significant interaction between experimental group and handicapping condition revealed that students classified as behaviorally disordered had differentially benefited on the math concepts subtest. Finally, a descriptive but non-significant difference favoring trained students was found on the math computation subtest.

The Effects of Coaching on the Standardized Test Performance of Mildly Handicapped Students

In recent years, researchers have attempted to identify sources of measurement error in handicapped populations. Such research is of importance because handicapped children are often among those most frequently tested in public schools, and because these populations have often been underrepresented in test standardization procedures (Fuchs, Fuchs, Dailey, & Power, 1985). Testing influences research has generally focused on the following issues: examiner effects, test anxiety and attitudes, and test-taking skills, or "test-wiseness" (Millman, Bishop, & Ebel, 1965).

Fuchs, Fuchs, Power, and Dailey (in press) tested handicapped (speech or language impaired) and nonhandicapped children using familiar and unfamiliar examiners and concluded that examiner familiarity had a differentially facilitating effect on handicapped children. This finding is supported by previous research efforts (Fuchs, Fuchs, Daily, & Power, 1985; Fuchs, Fuchs, Garwick, & Featherstone, 1983). Field (1981), however, found examiner familiarity or recent experience with nonhandicapped children had a negative effect on the test scores of developmentally handicapped preschool children. Dangel (1972) examined the influence of pretest referral information provided to examiners (examiner bias) on the intelligence scores of retarded students and reported that scores did not differ as a function of examiner bias.

Test anxiety and test attitudes have also been recently investigated with handicapped populations, but findings here have not always been in agreement. Bryan, Sonnefeld, and Grabowski (1983) reported that learning disabled (LD) students were more "test-anxious" than their nondisabled counterparts, while Sliwa (1977) found no such relation. Wolf (1975) reported that anxiety-reduction training had no effect on the performance of "test-anxious" behaviorally disordered (BD) boys. Finally, Scruggs, Mastropieri, Tolfa, and Jenkins (1985), and Tolfa and Scruggs (1985a) found that BD students reported more negative attitudes toward tests than their more average age peers.

In the area of test-taking skills, recent research has supported the notion that mildly handicapped (particularly LD) students exhibit deficiencies in this area with respect to standardized achievement tests. LD students have been shown to exhibit deficiencies in the use of prior knowledge and deductive reasoning strategies (Scruggs & Lifson, 1985), selection of appropriate strategies and attention to appropriate format features (Scruggs, Bennion, & Lifson, 1985; in press), and effective use of separate answer sheets (Tolfa, & Scruggs 1985b). Although standardized achievement tests have generally been found to be reliable and valid with mildly handicapped students (e.g., Pierce, 1984), results of the above test-taking skills research suggest that measurement error could be reduced (and consequently, scores improved) if mildly handicapped students could be successfully trained in "test-taking skills."

Much research has been conducted in the area of training in test-taking skills, but little of this research has addressed handicapped populations. In a recent meta-analysis, Scruggs, Bennion, and White (in press) examined the effects of such coaching on achievement test scores of elementary school children. They concluded that, in general, coaching had a very small overall effect on test scores, with somewhat larger effects being found for younger students, lower SES students, and students who had undergone longer training periods. No research was located in which mildly handicapped students had been trained, although, more recently, such training has been accomplished. Scruggs and Tolfa (1985) reported that a small sample of trained LD students had scored higher than controls on standardized word analysis test items, while no differences were found for reading comprehension items. These same findings were replicated by Scruggs and Mastropieri (in press) using a larger subject sample of LD and BD students. It was concluded that such training could have a strong facilitative effect (8-10 percentile points) on reading subtests with more complicated format demands, as suggested by Tolfa, Scruggs, and Bennion (in press). The findings of Scruggs and Mastropieri (in press) and Scruggs and Tolfa (1985), although encouraging, left several issues unaddressed. First, the subjects in these investigations were mostly primary level students generally less familiar with testing situations than older students. It would be of interest to know whether upper elementary students could benefit from such training. Second, training was only given in reading subtest areas, leaving open the question of whether such training could facilitate performance on mathematics subtests. Finally, only the Scruggs and Mastropieri (in press) investigation included BD students, and in that study, students were not stratified by handicapping condition and therefore analysis of any possible treatment by handicapping condition interaction was not possible. It was, therefore, the purpose of the present research to replicate and extend previous findings of training in test-taking skills to include (a) upper elementary students, (b) mathematics as well as reading subtests, and (c) separate analysis of test performance by different handicapping condition.

Method

Subjects

Subjects were 85 LD and BD students attending public schools in a western metropolitan area. Forty-four students had been classified as learning disabled and 41 students had been classified behaviorally disordered by national, state, and local standards. These standards included, for LD students, a forty percent discrepancy between ability (assessed by individual intelligence tests) and two areas of academic achievement. Although LD students in the present sample exhibited descrepancies in several different content areas, most had been referred for deficiencies in reading, followed by deficiences in mathematics functioning. Behaviorally disordered students were classified by teacher and school psychologist documentation of deficiencies in social or emotional functioning which interfered with classroom learning. These referrals were made for several different reasons, but in most cases students had exhibited aggressive or non-compliant behaviors in the classroom which interfered with routine classroom activities.

The sample included 21 4th, 38 5th, and 26 6th grade students, composed of 63 boys and 22 girls. Mean Weschler Intelligence Scale for Children-Revised for the experimental group was 92.45 (SD = 10.20). Mean WISC-R for the control groups was 91.48 (SD = 9.64). Achievement test scores for the sample are provided in the Results section.

Materials

Materials were developed specifically for the present investigation and consisted of (a) a practice test booklet with correct answers identified for practice with separate answer sheet, and (b) a practice test booklet with unmarked problems similar to, but not identical to, items in the Stanford

Achievement Test. Items were included which resembled those in two reading subtests (comprehension, word study skills) and three math subtests (concepts, computation, and word problems).

Procedure

Students were stratified by grade level and handicapping condition, and assigned at random to either a training or a no-treatment control condition. Training condition students were seen in small (1-6) groups by one of three trained experimenters for five 20-30 minute sessions. In the first session, students were given instruction and practice in the use of separate answer sheets using a practice test booklet for which correct items had been indicated with an arrow. Students were instructed in finding and monitoring their place on the answer sheet, marking and erasing carefully, and in checking their work. The second and third consisted of training in reading subtests. For the reading comprehension subtest, students were taught to refer back to the passage for recall questions, to use deductive reasoning strategies for inference questions, and to look for similarities between phrases or words in the passage and answer choices. For the word study skills subtest, students were taught to attend to appropriate cues and sound, rather than letter similarities in stem and option. For the math concepts subtests, students were taught to attend carefully to format changes. For the computation subtest, students were taught to carefully recopy problems on scratch paper in the most familiar form and more neatly. Finally, on the word problems subtest, students were taught to attend to command words in the problem and work problems carefully on separate paper. On all subtests, students were taught to (a) work quickly and carefully, (b) check answers if time permits, (c) answer all questions, (d) eliminate answers known to be incorrect, (e) incorporate prior or partial knowledge, and (f) become familiar with all subtest format demands.

The next week after training, all students were administered the Stanford Achievement Test by regular school personnel. Completed answer sheets were machine scored.

Results

Percentile scores were chosen for the present analysis because of their consistency across grade levels and because of their meaningfulness. Since previous research has indicated different effects are found for different subtests, separate condition (training vs. control) by handicap (LD vs. BD) analyses of variance (ANOVAS) were computed for each subtest. Significant differences were found for the word study skills and math concepts subtests, in favor of the training condition. On the word study skills subtest control students scored at an average of the 17.5th percentile, while trained students scored at an average of the 26.4th percentile, $F(1,81) = 4.79$, $p = .03$. No significant differences were found for handicapping condition, $F(1,81) = 1.53$,

$p < .56$ ($MS^e = 361.2$). On the math concepts subtest, control students scored at an average of the 16.4th percentile, while training condition students scored at an average of the 24.1st percentile, $F(1,81) = 4.54$, $p = .04$ ($MS^e = 288.3$). No significant difference was found for handicapping condition, $F(1,81) = 1.14$, $p = .29$, but an interaction effect was noted, $F(1,81) = 4.58$, $p = .04$, indicating differential facilitation on the part of the BD students. This interaction is depicted graphically as Figure 1.

Insert Figure 1 about here

Additionally, the main effect for experimental condition (but not handicap or interaction) approached significance on the mathematics computation subtest, $F(1,81) = 2.57$, $p = .11$. Descriptively, trained students scored at the 21.5th percentile while control students scored at the 15.5th percentile ($MS^e = 284.3$). Main effects or interactions did not approach significance on the reading comprehension or the math applications subtest (all $Fs < 1$). Descriptively, differences by condition were negligible, with experimental vs. control mean percentiles of 19.0 and 17.7, respectively, for reading comprehension; and 23.3 and 20.7 for math applications. In both cases, however, descriptive differences favored training condition students. Obtained effect sizes for all subtests are given in Table 1.

Insert Table 1 about here

Discussion

The findings of the present investigation replicate the findings of Scruggs and Tolfa (1985) and Scruggs and Mastropieri (in press) and extend them into upper elementary grades, mathematics subtests, and allow comparison of LD vs. BD student performance. That trained students outperformed controls on word study skills and mathematics concepts subtests supports the hypothesis of Tolfa, Scruggs, and Bennion (in press) that tests with more complicated formats may prove differentially difficult for mildly handicapped students. That is, the word study skills and mathematics concepts subtests each contain several potentially confusing format changes, while reading comprehension and math applications (i.e., word problems) subtests contain more "obvious" format demands, and fewer format changes. Although significant main effects were not found for total reading, total math, and total test, resulting effect sizes of these scores were substantially higher than those reported in the literature for nonhandicapped children (Scruggs, Bennion, & White, in press).

The obtained interaction by handicapping condition on the mathematics concepts subtest may simply represent characteristics of the present sample, but certainly deserves further research attention. Since mathematics functioning has been noted as a particular area of difficulty for BD students (Mastropieri, Jenkins, & Scruggs, in press), perhaps reflecting problems with attention and persistance of effort, it is possible that training in this case lessened the need to understand formats and thus represented a more valid indication of actual ability.

The results of this and previous research indicate that test-taking skills can be trained to mildly handicapped elementary age students, and that this training can significantly impact on test performance. Future research efforts are needed to assess whether similar training can also benefit secondary level mildly handicapped students, and whether training can improve scores on teacher-made tests. The present authors are currently investigating such possibilities (Taylor & Scruggs, 1983).

References

Bryan, J. H., Sonnefeld, L. J., & Grabowski, B. (1983). The relationship between fear of failure and learning disabilities. *Learning Disability Quarterly*, *6*, 217-222.

Dangel, H. L. (1972). Biasing effect of pretest referral information on WISC scores of mentally retarded children. *American Journal of Mental Deficiency*, *77*(3), 354-359.

Field, T. (1981). Ecological variables and examiner biases in assessing handicapped preschool children. *Journal of Pediatric Psychology*, *6*(2), 155-163.

Fuchs, D. (1985, April). *Exploring the norm in norm-referenced tests*. Paper presented at the annual meeting of the Council for Exceptional Children, Anaheim.

Fuchs, D., Fuchs, L. S., Dailey, A. M., & Power, M. H. (1985). The effect of examiners' personal familiarity and professional experience on handicapped children's test performance. *Journal of Educational Research*, *78*(3), 141-146.

Fuchs, D., Fuchs, L. S., Garwick, D. R., & Featherstone, N. (1983). Test performance of language-handicapped children with familiar and unfamiliar examiners. *The Journal of Psychology*, *114*, 37-46.

Fuchs, D., Fuchs, L. S., Power, M. H., & Dailey, A. M. (in press). Bias in the assessment of handicapped children. *American Educational Research Journal*.

Mastropieri, M A., Jenkins, V., & Scruggs, T. E. (in press). Academic and intellectual characteristics of behaviorally disordered children and youth. Monographs in Behavior Disorders, 9.

Millman, T., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness in the cognitive domain. Educational and Psychological Measurement, 18, 787-790.

Pierce, P. A. (1984). A comparative study of the California Achievement Test (Forms C and D) and the Key Math Diagnostic Arithmetic Test with secondary LH students. Journal of Learning Disabilities, 17, 392-296.

Scruggs, T. E., Bennion, K., & Lifson, S. (1985). An analysis of children's strategy use on reading achievement tests. Elementary School Journal, 85, 479-484.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press). Spontaneously produced test-taking skills of learning disabled students on reading achievement tests. Learning Disability Quarterly.

Scruggs, T. E. Bennion, K., White, K. (in press). Improving achievement test scores in the elementary grades by coaching: A meta-analysis. Elementary School Journal.

Scruggs, T. E., & Lifson, S. (1985, April). Are learning disabled students 'test-wise?' An inquiry into reading comprehension test items. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled children. Exceptional Children.

Scruggs, T. E., Mastropieri, M. A., Tolfa, D., & Jenkins, V. (1985). Attitudes of behaviorally disordered students toward tests. Perceptual and Motor Skills, 60, 467-470.

Scruggs, T. E., & Tolfa, D. (1985). Improving the test-taking skills of learning disabled students. Perceptual and Motor Skills, 6 , 847-850.

Sliwa, W. M. (1977). Self-esteem, general anxiety and test anxiety for learning disabled male students and for normal male students. Unpublished doctoral dissertation, Northern Illinois University.

Taylor, C., & Scruggs, T. E. (1983). Research in progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 50, 277.

Tolfa, D., & Scruggs, T. E. (1985a). Attitudes of behaviorally disordered students toward tests: A replication. Unpublished manuscript, Utah State University, Logan.

Tolfa, D., & Scruggs, T. E. (1985b). Can LD students effectively use separate answer sheets? Unpublished manuscript, Utah State University, Logan.

Tolfa, D., & Scruggs, T. E. & Bennion, K. (in press). Format changes in reading achievement tests: Implications for learning disabled students. *Psychology in the Schools.*

Wolf, A. D. (1975). *The effects of anxiety, grouping, and instruction on the performance of behaviorally disordered boys on a concept attainment task.* Unpublished doctoral dissertation, Columbia University Teachers College.

Author Notes

Preparation of this manuscript was supported in part by a grant from the United States Department of Education, Office of Special Education, Special Education Programs, #G008300008. The authors would like to thank Debra Peck for her assistance in the preparation of the training materials and Mary Ellen Heiner for typing the manuscript. We would also like to thank Dr. Joyce Barnes, Director of Special Education, and the teachers and students of Granite School District, Utah, for their cooperation and support of the research described here. Address requests for reprints to: Thomas E. Scruggs, UMC 68; Utah State University, 84322.

Table 1

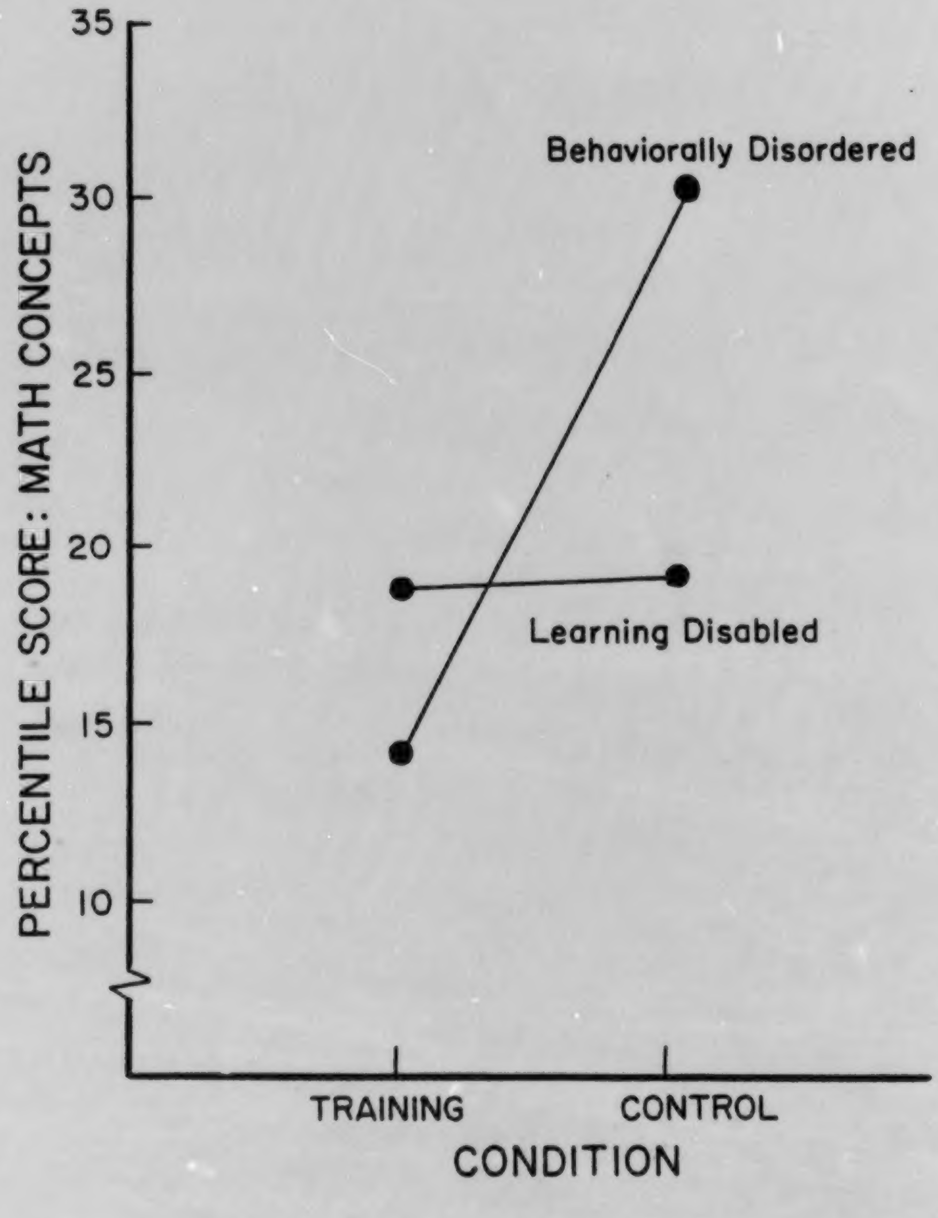
Obtained Effect Sizes

Subtest	Effect Size*
Reading Comprehension	.10
Word Study Skills	.53
Math Concepts	.59
Math Computation	.47
Math Applications	.15
Total Reading	.40
Total Math	.47
Total Test	.36

*All effect sizes were computed using control standard deviation as divisor and E-C mean differences in the numerator.

Figure Caption

Figure 1: Condition by handicap interaction: Math concepts.


PERCENTILE SCORE: MATH CONCEPTS
35
30
25
20
15
10
Behaviorally Disordered
Learning Disabled
TRAINING
CONTROL
CONDITION

APPENDIX P

Current Conceptions of Test-Wiseness:
Myths and Realities

Thomas E. Scruggs

Steven A. Lifson

Utah State University

Running head: MYTHS AND REALITIES

DISK 35; TOM/MANUS/6; MYTHS/REAL

Abstract

The popular conception of test-wiseness is reviewed and evaluated. Although some support for the concept of test-wiseness exists, in general the influence of test-wiseness with respect to: (a) contribution to measurement error, (b) cultural differences, (c) independence from general intelligence, and (d) facility for training, has been greatly overestimated. This paper attempts to place commonly found statements regarding test-wiseness in perspective of actual research findings.

Current Conceptions of Test-Wiseness:

Myths and Realities

It has been known for many years that all test scores reflect two additive elements: "true" score, accounting for the construct being measured, and "error" score (Magnusson, 1967). It has also been suggested that the error score may be itself composed of several additive components (Ebel & Damrin, 1960; Thorndike, 1951). These components have been said to include test anxiety (e.g., Sarason, 1978), achievement motivation (e.g., Atkinson, 1974; Chapman & Hill, 1971), and self-esteem (e.g., Roen, 1960). Such possible elements of measurement error have been discussed in detail by Jensen (1980).

Since 1965, an additional construct has been discussed repeatedly in the literature which is commonly thought to involve a substantial source of measurement error. This construct was defined by Millman, Bishop, and Ebel (1965), as "test-wiseness" (TW). Millman et al. defined TW as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score" (p. 707). They further described TW as "logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures" (p. 707). Ebel (1965) has suggested that error in measurement is more likely to be obtained from students low in test-wiseness. The student low in TW, therefore, may be more of a measurement problem than the student high in TW (Slakter, Koehler, & Hampton, 1970).

Analysis and Measurement of TW

Millman, Bishop, and Ebel (1965) have provided a definition and analysis of the construct on which most subsequent research has been based (Sarnacki, 1979). Millman et al. defined TW as distinct from general mental attitudes such as confidence and anxiety, and motivational states of the test-taker. In their analysis of TW, six elements were delineated. Four of these elements were considered to be independent of the test constructor or test purpose, while two were considered to be dependent on test constructor or test purpose. The four independent elements included (a) time using strategies, (b) error avoidance strategies, (c) guessing strategies, and (d) deductive reasoning strategies. Time using strategies included working quickly and efficiently and saving more difficult or time-consuming items for last. Error avoidance strategies included attending to directions, marking answers carefully, and checking all answers. Guessing strategies were considered to be the use of guessing when it was likely to benefit the test-taker. Deductive reasoning strategies included elimination of items known to be incorrect, item choices based on an analysis of the relation among items, such as choosing neither of two items which imply the correctness of each other (similar options), and use of content information from other test items and options.

The two elements thought to be dependent upon test constructor or purpose were intent consideration strategies and cue-using strategies. Intent consideration strategies included adopting the appropriate level of sophistication for the test, and considering the purpose of the test constructor. Cue-using strategies referred to the use of any consistent idiosyncrasies of the particular test constructor, such as inclusion of more true or false statements, placement of correct distractor, and grammatical inconsistencies between stem and options. Avoidance of items using the words "always" and "never" (specific determiners) was also considered a cue-using strategy.

Researchers have typically assessed TW in one of two indirect ways. One method is to teach TW skills to a population and assess the extent to which scores improve. The other method is to construct questions which are answerable only by use of specific TW skills and imbed these items in a larger test of answerable items. An example of an item answerable in terms of a TW strategy (similar options) was given by Slakter, Koehler, and Hampton (1970, p. 249):

"When Bestor cyrstals are added to water:

1. Heat is given off;
2. The temperature of the solution rises;

3. The solution turns blue;
4. The container becomes warmer."

The keyed answer to this item is (2), since the other options imply the correctness of each other. In a similar fashion, guessing strategies have been assessed by indicating a penalty for incorrect responses, and imbedding nonsense items for which no answer is correct. The extent to which subjects answer such nonsense items was considered a measure of guessing strategies (Slakter et al., 1970). Finally, such general TW strategies as use of prior or partial knowledge, deductive reasoning, and use of prior items have been assessed by administering reading comprehension test questions for which the referent reading passages have been deleted (e.g., Dunn, 1981; Scruggs & Lifson, 1985).

Since the initial analysis by Millman et al. (1965), a voluminous literature has emerged, reviews of which have been written by Bangert-Drowns, Kulik and Kulik (1983), Ford (1973), Fueyo (1977), Jones and Ligon (1981), and Sarnacki (1979). These reviews are all thorough to the extent that they cover adequately the body of literature referring to TW as it has been evaluated over the past two decades. It is the view of the present authors, however, that much of the influence associated with TW has been overstated to the point of distortion. It is the purpose of the present paper to clarify some issues regarding the construct "test-wiseness" and its consequences.

Commonly made statements regarding TW which are considered to be "myths" (by the present authors) include the following: (a) there is no substantial correlation between test-wiseness and intelligence, (b) TW constitutes a large source of variance which is commonly found in tests, (c) different American cultural groups are seen to differ substantially with respect to test-wiseness, and (d) test-wiseness is easily trained and results in substantial increases in test scores. These "myths" will be considered separately, followed by review of literature relevant to each, and a discussion of the realities associated with each particular myth.

Myth #1: TW is Not Substantially Related to General Intelligence

This myth is based largely upon the assumption that TW constitutes essentially an unfair advantage on test-taking tasks which some students have happened to acquire arbitrarily, while others have not. In addition, TW loses much credibility as a construct if it can be shown to be highly related to intelligence, and therefore not a specific, independent factor. Finally, if TW is not strongly related to intelligence, then it appears more likely that it can be easily trained; consequently, groups who can be shown to suffer with respect to TW would hypothetically benefit greatly from short instructional lessons in TW.

Millman et al. (1965) suggested that a test-wise subject would perform better on tests than would a less test-wise subject of equal intellectual ability. Wahlstrom and Boersma (1968) maintained, "while 'good' items may be used to control for error variance associated with test-wiseness, the writers contend that teacher-made achievement tests contain items with faults, and that test-wise subjects often received higher scores than subjects of equal intellectual ability" (p. 419).

The basis for this particular myth is found in a small number of empirical studies, whose interpretations have been greatly distorted. These investigations will be discussed in turn.

Dunn and Goldstein (1959) correlated scores on a group administered intelligence test (Army Aptitude Area 1) with scores on blocks of multiple choice items containing specific item flaws. These authors argued that since moderate correlations (.52-.72) were found between IQ and item blocks containing different TW cues as well as items containing no TW cues, "the ability to pick up cues on the type of material tested may be found at all levels of intelligence" (p. 178). In this investigation, however, no direct assessment of the relation between IQ and TW was made.

Kreit (1968) hypothesized that the intelligence of subjects is related to the acquisition of test-taking skills, and that more intelligent children would improve more from test session to test session. This hypothesis was not supported. Kreit reported only nonsignificant trends in the hypothesized direction. In this investigation, however, narrow and overlapping groups comprising his sample precluded a fair assessment of his hypothesis. This author, then, did not demonstrate the lack of a strong relation, but merely failed to support his own predictions with respect to one aspect of the TW/intelligence issue.

The most commonly cited study with respect to test-wiseness and intelligence was conducted by Diamond and Evans (1972). These researchers concluded that TW is cue-specific (that is, not one general ability) and that the overall correlation between the aspects of TW tested was not strong. In fact, the overall correlation between IQ and TW reported by Diamond and Evans was .49 which, if corrected for attenuation of the somewhat unreliable test-wiseness test, becomes a correlation of .61. In either case, the obtained correlation is strong enough to constitute a moderate relation between test-wiseness as measured and general ability. The conclusions of Diamond and Evans, although unwarranted, have been consistently cited by others more interested in perpetuating the myth of this aspect of TW than accurately reporting the data.

Other researchers, not as widely cited, have provided stronger information that TW and intelligence are in fact related. Anderson (1973) reports, "analysis of the correlational data indicates that for the total sample a significant (though moderate) correlation is obtained between TW and mental ability, between TW and achievement, and betwen TW and deductive reasoning ability" (page 89). Millikin (1975) correlated performance on a test-wiseness test and a general mental ability test on a sample of 306 eleventh grade subjects, and found a significant relation between a measure of general ability and TW.

Taken as a whole, the bulk of the research literature seems to indicate that a substantive correlation is typically found between TW and tests of mental ability, allowing for a tangible amount of shared variance. Apparently, however, these findings have not satisfied other authors in the field of TW, for the above articles are generally selectively cited as providing evidence that TW and intelligence are not correlated significantly. Thus, Dillard, Warrior-Benjamin, and Perrin (1977) maintained, "Kreit (1967) found that improved test-wiseness and intelligence were not significantly related" (p. 1135). Likewise, Crehan, Gross, and Koehler (1978) cited Diamond and Evans and reported, "previous research has shown that TW is not highly related to cognitive ability" (p. 40). Crehan, Koehler, and Slakter (1974) also cited Diamond and Evans and reported, "investigators examining the cognitive correlates of TW have concluded that TW is not highly related to cognitive ability" (p. 209). This myth has also been maintained by those who simply assert that students equal in intelligence may differ in TW. For instance, Gross (1977) asserted "(TW) concerns the extent to which examinees of similar ability or achievement received different test scores as a result of differences in test-taking shrewdness" (p. 97). Wahlstrom and

Boersma (1968) asserted ". . . test-wise Ss often receive higher scores than Ss of equal intellectual ability" (p. 419).

It can, therefore, be seen that in spite of substantial evidence linking general reasoning ability and measures of test-wiseness, researchers have continued to report the lack of a relation between the two variables. The reasoning for this is uncertain, although it no doubt reflects in part an interest in (a) defending the construct of TW as one separate from intelligence, and (b) consequently, implying that such ability is easily trained and manipulated. To this end, relevant data have been misinterpreted, or simply ignored. In addition to the empirical findings of correlations between TW and intelligence, and the methodological errors of those who maintain there is no such relation, an appeal to "common sense" can be made. High on the list of Millman, Bishop, and Ebel's analysis of test-wiseness is what is referred to as "deductive reasoning strategies", of which are included elimination of options known to be incorrect, elimination of options which imply the correctness (or incorrectness) of each other, utilization of relevant content information in other test items, and choice of items which encompass all of two or more given statements known to be correct. Other strategies include a deduction of the intent of the test constructor and a determination of regularities in stem or option cues on the part of the test constructor. It would defy credibility to assert that these "deductive reasoning" strategies are not related to general mental ability.

As with most myths, however, elements of truth remain. If it is obvious that many test-taking strategies are strongly dependent upon the reasoning skills of the test-taker, it is also obvious that some other strategies can be easily taught and involve little reasoning ability. These include such strategies as working quickly, moving past items which resist a quick response, answering all questions, using time remaining after the completion of tests to reconsider answers, asking the examiner for clarification of ambiguous questions, guessing whenever necessary, and developing prior familiarity with specific test format demands. These strategies also comprise a component of test-wiseness and have been successfully trained to mildly handicapped students at the primary-age level, to the extent that performance on achievement tests has been enhanced (Scruggs, 1984a, b; Scruggs & Mastropieri, in press b). Although such strategies as those previously mentioned do not typically appear on tests of "test-wiseness," these strategies may be, in fact, somewhat independent of intelligence and therefore subject to relatively simple remediation. To this extent, then, the issue of test-wiseness not being related to intelligence does have some support. To the extent to which this myth has been reported in the literature, however, it must be challenged--that is, TW is not a construct which students happen to acquire by chance or serendipity, which is unrelated to intelligence, and which results in substantial fluctuations of scores in achievement tests.

Myth #2: TW Constitutes a Large Source of Variance/TW Cues are Commonly Found on Tests

Although it is clear that some students are less able to "outguess" certain test items than their "test-wise" peers, the issue at stake in this particular myth revolves around whether or not the amount of variance associated with TW is large. Some authors have simply reported that TW is a potential source of error. Gross (1977) argues, "Millman, Bishop, and Ebel (1965) have advocated that TW be taught to minimize inter-examinee TW differences, thereby reducing measurement error. . ." (p. 97). Gross (1977), referring to Ebel (1965), writes, "more error in measurement is likely to originate from students who have too little, rather than too much, skill in taking tests" (p. 97). Sarnacki (1979) writes, "TW is widely recognized as a source of additional variance in test scores and is a possible depressor of test validity" (p. 253). Some authors, however, have magnified the importance of this argument and have written that, in fact, the source of error in test-wiseness is extensive. Thus, Wahlstrom and Boersma (1968) maintained, "an important source of variation in test scores is test-wiseness" (p. 413). McPhail (1978) argued, "test-wiseness operates as error variance and its effect is to reduce the validity and reliability of tests" (p. 168). Kalechstein, Kalechstein, and Doctor (1981) maintained, "test-wiseness has been considered a potentially large source of error variance" (p. 198).

The fact that TW accounts for a source of error variance is indisputable. The question here is whether, in fact, TW constitutes a large source of variance and whether TW cues are commonly found in tests. The basis for the magnitude of the effect of TW derives largely from a confusion between the terms "statistically significant" and "practically important." For example, Sarnacki (1979) cites a number of studies for which statistically significant increases in test scores were associated with training in TW (e.g., Callenbach, 1973; Gross, 1976; Oakland, 1972). Although Sarnacki is correct that these researchers did, in fact, exert a "significant" increase in test scores as a result of training in TW, the fact is that in virtually all cases, the effect sizes were quite small (this issue will be discussed further under the "easily trained" myth). In fact, the very studies that Sarnacki cites are stronger arguments in favor of the issue that TW is a relatively small source of variance in achievement test scores. One specific study is worthy of mention. Sarnacki cites Gross (1976) as evidence that significant increases in test scores were associated with training in TW. A review of this dissertation, however, demonstrates that three selected TW behaviors were taught. These behaviors included risk taking, deductive reasoning, and time using. The dependent measure was the Metropolitan Achievement Test (MAT) Advanced Battery. Gross concluded that (a) deductive reasoning was not successfully taught (see "TW not correlated with IQ" myth), (b) risk taking (i.e., guessing) exerted a significant influence on test score only when guessing was inhibited in control conditions, and (c) although time using was successfully taught, it did not affect test score. Thus, the very dissertation cited by Sarnacki suggests that TW constitutes a relatively small source of variance.

In one of the most thoughtful investigations of TW, Rowley (1974) administered vocabulary and mathematics test items in both free response and multiple choice formats. Partial correlations were computed between scores on multiple choice items and measures of TW and risk-taking (RT), with free response scores partialed out. Rowley found significant partial correlations between vocabulary scores and TW and RT measures, and concluded that use of multiple choice tests "can result in high risk-taking, test-wise examinees scoring more highly than other examinees whose knowledge and ability are the equal of theirs" (p. 21). Analysis of the actual extent of performance advantage of students high in TW is difficult, because gain scores (from free response to multiple choice) were not reported. Examination of correlational data, however, indicates that TW and RT were not correlated at all with mathematics multiple choice items (partial r's = near 0) and that the partial correlations with vocabulary items were not high (r's of .27 and .14 for TW and RT, respectively) when guessing was not penalized (see Gross, 1976). In this investigation, then, TW was seen to account for 7% of the variance in vocabulary test performance, while RT accounted for less than 2% of total vocabulary test variance. When this finding is considered with the near zero correlations between TW, RT, and mathematics test performance, the conclusion that such factors constitute a large source of variance is difficult to justify.

Another argument in favor of the "large source of variance" myth comes from analyses of tests themselves. Metfessel and Sax (1958) looked for bias in placement of key to correct answers and found that more questions were keyed "true" on true-false tests than "false." They argued that 42% of the tests that they studied were found to have answer placement flaws that may conspire with response sets to artificially inflate scores. Even if these data are true, the point remains that test-takers would need to know ahead of time in which direction keyed items were biased in order to make any benefit of these flaws. The strongest argument with respect to Metfessel and Sax' analysis, however, is that although they document the possibility of placement flaws which may artifically inflate scores, they offer no quantitative data which support that these cues actually do result in inflated scores.

In order to investigate more fully whether TW cues are commonly present in achievement tests, the present authors have recently examined five major standardized achievement tests (California Achievement Test, Metropolitan Achievement Test, Comprehensive Test of Basic Skills, Iowa Test of Basic Skills, and Stanford Achievement Tests) for presence of TW cues, including specific determiners, similar options, stem options, or absurd options as defined by Slakter et al. (1970). We independently evaluated all test items for the presence of these cues and afterwards computed a 96% coefficient of agreement on TW cues. Nevertheless, we found that such TW cues exist in less than half of 1% of items on all these tests, substantially different from the "large source of variance" TW cues are supposed to encompass.

Another argument which can be made is that although such cues are not commonly present in standardized tests, they are present to a large extent in teacher-made tests. To this end, some studies have indicated that training in TW skills does not critically influence performance on standardized achievement tests but does influence performance on multiple-choice tests with poorly made distractors, which are then argued to be representative of teacher-made tests. Thus, Wahlstrom and Boersma (1968) have argued that TW training increases scores on "poorly made" tests but does not increase scores on standardized test items. Although there may or may not be some truth to this argument, there is a logical flaw in it. Those who advocate training in TW to improve scores on poorly constructed test items are in essence arguing that teachers should teach their students how to outguess their poorly constructed tests. Such an argument is not logically sensible, and in addition, suggests that outguessing test items for which the content is not known would result in more, rather than less, measurement error. At any rate, the interests of the teacher and students would be better served by putting additional time into training the teacher to construct better items, rather than teaching the students to outguess them more effectively.

Myth #3: Cultural Differences Exist in TW

It has been assumed as far back as the "codification" of TW in the original article by Millman, Bishop, and Ebel (1965) that TW of the type found on objective tests is culturally determined. One of the more widely cited references to this myth is by Millman and Setijadi (1966) who compared the performance of American and Indonesian students on open-ended and multiple-choice questions. The American students enjoyed an advantage on the objective questions, even after the Indonesian students were familiarized with the mechanics of choosing the correct answer. Furthermore, Lo and Slakter (1973) compared Chinese and American students on an instrument meant to measure TW and risk-taking in test circumstances. These two articles have been commonly cited by researchers as evidence that some ethnic/cultural groups in the United States may score lower on achievement tests because of "cultural" differences in TW. This possibility has led to much research on training American minority groups on TW skills. Often, however, deficiencies in TW exhibited by minority groups have simply been assumed rather than documented. Slakter, Kohler, and Hampton (1970) maintain "the objectives of [a TW] learning program would be not only to decrease the errors of measurement mentioned by Ebel (1965, p. 206) but to decrease the handicap under which many examinees apparently operate. For example, certain subsets of the population (black students, rural students, etc.) score lower on achievement tests than the population at large" (p. 253). The assumption by these authors is that much of the difference in achievement test scores is due to cultural influences in TW, and not lower levels of achievement in general.

Evidence presented to support the assertion that minority groups lack TW, however, is often tenuous. For example, when Kalechstein, Kalechstein, and Doctor (1981) cited Ortar (1960), among others, in their statement, "several investigators have noted the lack of test-wiseness in culturally different children" (p. 198), they implicitly referred to American minorities. Ortar actually speaks of the difficulties in using standardized tests when faced with a culturally diverse population, stating that under such circumstances, the assumption of equality of past experience cannot be made. It is not clear that this statement is accurate when applied to inner city, black, or lower socioeconomic status students.

Most empirical studies attempting to document differences in TW between ethnic/cultural groups consist of either (a) the administration of a TW instrument to different cultural groups, or (b) attempts to evaluate the impact of TW training on the subsequent scores on a TW instrument or a real standardized test. Despite the concern expressed by many researchers (e.g., Ebel, 1965; Ortar, 1960) that score differentials may be related to between-group deficits in TW, relatively little research has focused on identifying that deficit. For example, Kalechstein et al. (1981) cited previous investigators who have described the lack of TW in culturally different/disadvantaged groups, but themselves administered a TW training program to a group of black, disadvantaged second graders without reference to a supposedly "advantaged" group. However, it may be that all second graders as a group are relatively inexperienced with tests in general. The performance of black second graders after exposure to a TW treatment in the absence of comparison to other groups, therefore, tells us relatively little concerning cultural group differences in TW. Thus, Kalechstein et al. have not established that achievement tests are less valid for the group they studied. What they have done is replicated the study by Callenbach (1970) with a different population and raised questions not directly addressed in their own investigation. Likewise, Dreisbach and Keogh (1982) successfully trained TW skills to Mexican-American children and commented "test-wiseness may be particularly important when testing children from economically disadvantaged backgrounds and/or where the primary language of the home is not standard English" (p. 228). Although language of test administration and language competence of the child were also investigated, the primary focus of this investigation was the hypothesis that Mexican-American children "lack 'test-wiseness' and thus do poorly on tests" (p. 224). Differential effects of training for low SES or minority populations, however, were not investigated in their study and leave unanswered the issue of whether such training is in fact "particularly important" for low SES or minority populations.

In contrast to the questionable support of cultural/minority differences in TW, there is evidence that these groups differ little with respect to TW. In a dissertation by Yearby (1975) in which SES, race, and sex were controlled, no significant differences were observed between the groups on the test-taking skills pretest. Another study which directly addressed the question of whether disadvantaged or minority populations lack TW was conducted by Diamond, Ayres, Fishman, and Green (1976). Although the study was clearly designed to indicate relative deficiencies in TW on the part of black inner-city children, support for this hypothesis was not found. It was found that black inner-city children performed significantly above chance on a TW instrument, and that scores on the TW instrument did not predict grades on the Verbal Achievement subtest of the California Achievement Test. This suggests that it can neither be assumed that disadvantaged or minority groups lack TW, nor that a relation between TW and achievement test scores exists in these groups. In a review by McPhail (1976), it was concluded that "TW studies conducted on black and other minority student populations . . . have been inconclusive" (p. 168). Although it may be argued that direct evaluations of relative levels of test-wiseness in minority and nonminority groups are lacking, it must be maintained that at present the assertion of American minority groups being lower in test-wiseness, and this deficiency being responsible for much of the performance differences between groups, is largely unsupported.

As in most contemporary myths, however, a degree of truth can be discerned. Although studies which compare the effectiveness of test-wiseness training between minority and nonminority groups have not been found, a recent investigation does offer some support for the "cultural difference in TW" issue. Through meta-analysis procedures, Scruggs, Bennion, and White (in press) have been able to make quantitative comparisons in the effectiveness of

TW training on achievement test scores of minority and nonminority groups which were not directly assessed by individual studies. Scruggs et al. evaluated 24 empirical studies which investigated the effects of TW training on elementary school students, grades 1 through 6. It was found that with less than 4 hours of treatment, neither "low SES" nor "not low SES" subjects benefited appreciably (average effect sizes of -.05 and .08). With more than 4 hours of treatment, students from low socioeconomic background benefited more than twice as much as students who were not from low SES backgrounds (average effect sizes of .44 vs. .20). Since low SES subjects under these circumstances appeared to benefit more than twice as much as their counterparts from higher SES groups, the finding implies that children from low SES backgrounds are somewhat deficient with respect to TW. In addition, most students representing low SES groups in the studies evaluated were also members of inner city minority groups. It must be noted, however, that the effect size differential for a student receiving 4 or more hours of treatment from low SES and not low SES backgrounds was .24 standard deviation units, a relatively small difference which in no way could account for the large performance differences seen between SES groups on achievement tests. Although the Scruggs et al. (1984) study provides some evidence that students from low SES and minority backgrounds may suffer somewhat with respect to TW skills, these deficiencies explain little of performance differences between the two groups.

Myth #4: TW Is Easily Trained and Results in Large Gains in Test Performance

This myth is related to the "large source of variance/ commonly found" myth in which statistical significance has been confused with practical importance. For example, Sarnacki (1979) referred to Gaines and Jongsma as having concluded "that TW can be taught in a relatively short amount of time with significantly higher performance on standardized tests resulting." Slakter goes on to cite several others who "significantly" raised achievement test scores by TW training (e.g., Callenbach, 1973; Gross, 1976; Wahlstrom & Boersma, 1968). An analysis of a number of significant versus nonsignificant differences, however, says little about the relative size of the effect of training. In a recent meta-analysis, Bangert-Drowns, Kulik, and Kulik (1983) indicated that training in TW resulted in average effect sizes on achievement test scores of .29. On the primary grade levels, this effect size would be equivalent to approximately three months of academic achievement, not a large difference by educational standards. In a more recent meta-analysis, however, using somewhat different criteria for evaluating effect sizes, Scruggs, Bennion, and White (in press) determined that the average effect size in the elementary grades for raising scores on achievement tests was .10, less than half of that reported by Bangert-Drowns et al., reflecting grade equivalent increases of questionable significance. It was only after relatively long-term training (i.e., longer than four hours) that the resulting effect sizes began to resemble those reported by Bangert-Drowns et al. This finding demonstrated by meta-analysis in the elementary grade level has recently been demonstrated to be true with college-bound students on the Scholastic Aptitude Test (De Simonian & Laird, 1983). Thus, it appears that the notion that TW is easily trained and results in substantially higher test scores is unjustified.

Another argument that TW is easily trained comes from researchers who trained selected aspects of TW and measured performance on the basis of a TW instrument (e.g., Gibb, 1964; Slakter et al., 1970; Moreshultz & Baker, 1966). It was found that TW training does substantially and easily increase scores on TW instruments, and these findings have been supported by the meta-analysis of Scruggs et al. (in press). Although this type of training does seem to be effective in promoting scores on TW tests, the extent to which this training raises scores on actual tests remains relatively small. Another argument offered by those who maintain TW is "easily trained" is that although TW cues are not common on standardized achievement tests, they are common on flawed teacher-made tests, and it is on these types of tests that TW training is most beneficial. This issue has been addressed above. Although it seems absurd for teachers to teach their students to "outguess" their own poorly constructed tests, the idea of training teachers to construct better test items is often dismissed out of hand. Sarnacki (1979) argues unconvincingly that even if teachers are trained in the principles of TW, item faults may still occur. One may just as easily assert that students may forget some of the TW skills they were taught. In fact, if the same amount of time was spent training teachers to construct better test items, it is logical to assume that less, rather than more, error would result than if students were trained to guess correctly the answers to questions they do not understand.

In summary, it can be stated that (a) relatively small gains in standardized test performance have been achieved only after extensive training, and (b) although effects are greater for poorly constructed items, training in this area is more difficult to justify.

In spite of this present, rather pessimistic appraisal of the "easily trained" myth, however, a positive hypothesis, which has only recently received some research support, does remain. Although group differences with respect to TW training have been relatively small, it is possible that there exist certain individuals (or small groups) for whom TW is both necessary and beneficial and for whom relatively large differences in performance can be achieved. It has been seen that students classified as mildly handicapped (i.e., learning disabled and behaviorally disordered) may differ from their nonhandicapped peers with respect to (a) attitudes toward tests (Scruggs & Mastropieri, in press a), and (b) spontaneous production of effective test-taking strategies, including the effective utilization of test format (Scruggs, Bennion, & Lifson, in press a), selection of an appropriate test-taking strategy (Scruggs, Bennion, & Lifson, in press b), and use of prior or partial knowledge and deductive reasoning (Scruggs & Lifson, 1985). A recent experiment in TW training of regular third grade students has indicated that TW training benefited the lower half of the class much more so than the upper half (Scruggs, Bennion, & Williams, 1984). Such differences were seen to "wash out" when scores of the trained group as a whole were combined. Finally, successful training of test-taking skills has recently been achieved in special education populations (Dunn, 1981; Lee & Alley, 1981; Scruggs, 1984; Scruggs & Mastropieri, in press b). The obtained effect sizes in these initial investigations have tended to be somewhat larger than those obtained on nondisabled populations, and there is the added feature that many of these students are functioning within a level at which relatively slight changes for better or worse on achievement test performance may result in more serious decisions regarding educational placement. In other words, although gains have typically been small and of less consequence for normally achieving students, even relatively small gains may be of greater importance to students functioning at the lower end of the distribution. Also, mildly handicapped groups do in fact exhibit less efficient test-taking strategies than their nonhandicapped peers, and it would seem logical to assert that these students should be trained to utilize the same strategies that other students are spontaneously using.

Summary and Conclusions

The present view has attempted to critically evaluate four contemporary myths associated with test-wiseness. In this article, we have stated that (a) the disassociation of TW from general cognitive ability has not been verified, (b) TW has not been shown to constitute a large source of error variance in tests, (c) American minority groups have not been shown to be seriously lacking in TW, and (d) relatively modest improvement in test scores has been achieved only through long and intensive training in TW skills. Stated more positively, TW can be said to be a tangible component of the test-taking experience but one which nevertheless plays a relatively minor role in overall test scores for most students.

Several implications can be drawn from this analysis for the practicing school psychologist. First, in many individual cases, it may be wiser to assume TW has played a relatively minor role in test performance. Although teachers often explain a particular student's poor test scores by asserting he/she is simply a poor "test-taker," such reports may reflect either a well-intentioned but misguided sympathy for the student, or simply a misreading of the student's actual abilities. A psychologist who has been told that a particular child's low scores reflect only poor test-taking skills would be well advised to seek more tangible evidence that this is truly the case. Second, if it can be demonstrated that a given student is exceptionally weak in TW, there is little reason to believe that that student could not be trained in TW skills. Finally, in the case of special education students, it may be advisable to ensure that all such students have had some additional guided practice on unfamiliar test formats.

It can be concluded that although TW as a construct is weaker and less pervasive than commonly assumed, there is nevertheless tangible evidence of its (perhaps multifaceted) existence and some indication that, although large groups tend to gain little from specific training in TW, there may be certain individuals or smaller groups for whom the construct of TW does constitute an "important source of error." Further research in this area may do much to ultimately clarify the issue of test-wiseness.

References

Anderson, B. E. (1974). The effects of test-wiseness upon mental ability measurements, achievement and deductive reasoning in a college sample: An instructional model (Doctoral dissertation, the University of Connecticut, 1973). Dissertation Abstracts International, 34, 3975-A.

Atkinson, J. W. (1974). Motivational determinants of intellective performance and cumulative achievement. In J. W. Atkinson, J. O. Raynor et al., Motivation and achievement (Ch. 20). Washington, D.C.: Winston.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. (1983). Effects of coaching programs on achievement test performance. Review of Educational Research, 53(4), 571-585.

Callenbach, C. A. (1972). The effects of instruction and practice in nonsubstantive test-taking techniques upon the standardized reading test scores of selected second grade students. (Doctoral dissertation, Pennsylvania State University, 1971). (University Microfilms International Order Number 72-13, 826)

Callenbach, C. A. (1973). The effects of instruction and practice upon the standardized reading test scores of selected second grade students. Journal of Educational Measurement, 10(1), 25-30.

Chapman, M., & Hill, R. A. (Eds.) (1971). Achievement motivation: An analysis of the literature. Philadelphia: Research for Better Schools, Inc.

Crehan, K. D., Gross, L. J., Koehler, R. A., & Slakter, M. J. (1978). Developmental aspects of test-wiseness. Educational Research Quarterly, 3(1), 40-44.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. (1974). Longitudinal studies of test-wiseness. Journal of Educational Measurement, 11(2), 209-212.

De Simonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. Harvard Educational Review, 53(1), 1-15.

Diamond, J. J., Ayrer, J., Fishman, R., & Green, P. (1976). Are inner city children test-wise? Journal of Educational Measurement, 14, 39-45.

Diamond, J. J., & Evans, W. (1972). An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 9, 145-150.

Dillard, M., Warrior-Benjamin, J., & Perrin, D. W. (1977). Efficacy of test-wiseness on test anxiety and reading achievement among black youth. Psychological Reports, 41, 1135-1440.

Dreisbach, M., & Keogh, B. K. (1982). Testwiseness as a factor in readiness test performance of young Mexican-American children. Journal of Educational Psychology, 74, 224-229.

Dunn, A. E., Jr. (1981). An investigation of the effects of teaching test-taking skills to secondary learning disabled students in the Montgomery County (Maryland) Public Schools Learning Centers (Doctoral dissertation, George Washington University).

Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 19, 171-179.

Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L., & Damrin, D. E. (1960). Tests and examinations. In C. W. Harris (Ed.), Encyclopedia of educational research (3rd ed.). New York: Macmillan Co.

Ford, V. A. (1973). Everything you wanted to know about test-wiseness. (ERIC Document Reproduction Service No. ED 093 912)

Ford, V. A. (1976). The influence of two test-wiseness programs upon students' test performance (Doctoral dissertation, Pennsylvania State University). Dissertation Abstracts International, 37, 7037A.

Fueyo, V. (1977). Training test-taking skills: A critical analysis. Psychology in the Schools, 14, 180-185.

Gross, L. J. (1976). The effects of three selected aspects of test-wiseness on the standardized test performance of eighth grade students. (Doctoral dissertation, State University of New York at Buffalo, 1975). *Dissertation Abstracts International*, *36*, 6551A.

Gross, L. J. (1977). The effects of test-wiseness on standardized test performance. *Scandinavian Journal of Educational Research*, *21*(2), 97-111.

Hecht, J. T. (1973). Test-wiseness and usability of scores obtained from repeated IQ test administrations (Doctoral dissertation, Southern Illinois University). *Dissertation Abstracts International*, *33*, 4937A-4935A.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jones, P., & Ligon, G. D. (1981). *Preparing students for standardized testing: A literature review*. Austin, TX: Austin Independent School District.

Kalechstein, P., Kalechstein, M., & Doctor, R. (1981). The effects of instruction on test-taking skills in second grade black children. *Measurement and Evaluation in Guidance*, *13*(4), 198-202.

Kreit, L. H. (1968). The effects of test-taking practice on pupil test performance. *American Educational Research Journal*, *5*, 616-625.

Lee, P., & Alley, G. R. (1981). Training junior high school LD students to use a test-taking strategy. Lawrence, KS: Kansas University. (ERIC Document Reproduction Service No. ED 217 649).

Lifson, S., Scruggs, T. E., & Bennion, K. (1984). Passage independence in reading achievement tests: A follow-up. Perceptual and Motor Skills, 58, 945-946.

Lo, M. Y., & Slakter, M. J. (1973). Risk taking and test-wiseness of Chinese students. The Journal of Experimental Education, 42(2), 56-59.

Magnusson, D. (1967). Test theory. Reading, MA: Addison-Wesley Publishing Co.

McPhail, I. P. (1978). A psycholinguistic approach to training urban high school students in test-taking strategies. The Journal of Negro Education, 47(2), 168-176.

Metfessal, N. S., & Sax, G. (1958). Systematic biases in the keying of correct responses on certain standardized tests. Educational Psychological Measurement, 18, 787-790.

Millikin, J. L. (1975). Some correlates of test-wiseness among high school students. (Doctoral dissertation, Texas A & M University). Dissertation Abstracts International.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. Educational and Psychological Measurement, 25, 707-726.

Millman, J., & Setijadi (1966). A comparison of the performance of American and Indonesian students on three types of test items. The Journal of Educational Research, 59, 315-519.

Oakland, T. (1972). The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. Journal of School Psychology, 10, 355-360.

Ortar, G. (1960). Improving test validity by coaching. Educational Research, 2, 137-142.

Roen, S. R. (1960). Personality and Negro-white intelligence. Journal of Abnormal and Social Psychology, 61, 148-150.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple choice tests? Journal of Educational Measurement, 11, 15-23.

Sarason, I. G. (1978). The Test Anxiety Scale: Concepts and research. In C. D. Spielberger & I. G. Sarason (Eds.), Stress and anxiety (Vol. 5). Washington, D.C.: Hemisphere.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49, 252-279.

Scruggs, T. E. (in press). The administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school children. Final Report. (ERIC Document Reproduction Service)

Scruggs, T. E. (1984). Improving the test-taking skills of learning disabled students. Unpublished manuscript, Utah State University, Logan, Utah.

Scruggs, T. E., Bennion, K., & Lifson, S. A. (in press a). An analysis of children's strategy use on reading achievement tests. Elementary School Journal.

Scruggs, T. E., Bennion, K., & Lifson, S. A. (in press b). Spontaneously employed test-taking skills of learning disabled students on reading achievement tests. Learning Disabilty Quarterly.

Scruggs, T. E., Bennion, K., & White, K. R. (in press). Teaching test-taking skills to elementary grade students: A meta-analysis. In T. E. Scruggs, The administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school children. Final Report. (ERIC Document Reproduction Service)

Scruggs, T. E., Bennion, K., & Williams, N. J. (1984). Effects of training in test-taking skills on test performance, attitudes, and on-task behavior of elementary school children. Unpublished manuscript, Utah State University, Logan, UT.

Scruggs, T. E., & Lifson, S. A. (1985, April). Are learning disabled students 'test-wise?' An inquiry into reading comprehension test items. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Scruggs, T. E., & Mastropieri, M. A. (in press a). Attitudes of behaviorally disordered students toward tests. *Perceptual and Motor Skills*.

Scruggs, T. E., & Mastropieri, M. A. (in press b). Training test-taking skills to behaviorally disordered and learning disabled students. *Exceptional Children*.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970). Learning test-wiseness by programmed texts. *Journal of Educational Measurement*, *7*, 247-254.

Thorndike, R. L. (1951). Community variables as predictors of intelligence and academic achievement. *Journal of Educational Psychology*, *42*, 321-338.

Wahlstrom, M., & Boersma, F. J. (1968). The influence of test-wiseness upon achievement. *Educational and Psychological Measurement*, *28*, 413-420.

Yearby, M. E. (1976). The effect of instruction in test-taking skills on the standardized reading test scores of white and black third-grade children of high and low socioeconomic status. (Doctoral disseration, Indiana University, 1975). *Dissertation Abstracts International*, *36*, 4426-A.

Author Notes

The preparation of this manuscript was supported in part by a grant from the U.S. Department of Education, Research in the Education of the Handicapped, #G008300008. The authors would like to thank Mary Ellen Heiner and Robert LaMont for their assistance in the preparation of this manuscript.

APPENDIX Q

Academic and Intellectual Characteristics of Behaviorally Disordered Children and Youth

Margo A. Mastropieri, Vesna Jenkins, and Thomas E. Scruggs

Utah State University

Running head: ACADEMICS

Abstract

Research describing academic and intellectual characteristics of behaviorally disordered (BD) students is reviewed. Investigations reviewed in this paper have focused on areas of intellectual, academic, and psycho-social functioning as they pertain to school achievement. In general, it has been found that BD students exhibit academic deficiencies greater than those exhibited on tests of intellectual functioning and perform below average in all content areas, with particular discrepancies noted in math functioning. In addition, variables such as locus of control, responses to the test-taking situation, and attitudes toward academic tasks, may covary with academic performance.

Academic and Intellectual Characteristics of Behaviorally Disordered Children and Youth

All students classified as behaviorally disordered (BD) by definition are in need of programming designed to improve social or emotional functioning. Since most of this programming occurs in academic environments, however, it is important to know whether students so classified also exhibit deficiencies with respect to intellectual or academic functioning. If BD students are generally found to be deficient in academic functioning, it may be necessary to incorporate remedial instruction as a major component of the educational environment. This review is intended to synthesize academic and intellectual characteristics of behaviorally disordered children and youth in order to provide a basis for future research and practice.

Two data bases (Psychological Abstracts, ERIC) were examined for data-based articles pertaining to academic and intellectual characteristics of BD students. In addition, recent books on behavior disorders (e.g., Kauffman, 1985) were reviewed for sources. Finally, past issues of the journal Behavioral Disorders and the series Monographs in Severe Behavior Disorders of Children and Youth were examined for relevant articles. Articles were included which selected a population on the basis of disturbances in social or emotional functioning, exclusive of psychotic or autistic samples. By these means, 25 articles reporting data were located and are given in Table 1.

Insert Table 1 about here

The investigations reviewed here represent a wide range of samples of children and youths referred to as "behaviorally disordered." To this extent, any general agreement between investigations suggests broad generalizability. When research reports disagree, however, interpretations are more difficult. In general, descriptions of academic and intellectual characteristics can be divided into three main areas: (a) intelligence, (b) achievement, and (c) psycho-social functioning and academic performance.

Intelligence

Studies of intellectual functioning are of relevance to the study of academic characteristics for two reasons: (a) IQ consistently has been a strong predictor of academic achievement (Kauffman, 1985), and (b) IQ scores can provide information concerning ability/achievement discrepancies. The following section describes the results of several investigations of intellectual performance.

In 1964, Stone and Rowley reported a mean IQ of 96.5 (ranging from 62 to 135) for 116 children referred for psychiatric services. Graubard (1964) found 21 delinquent or neglected boys in psychiatric residential treatment for two to eight years to have a mean IQ of 92.3 (range 71 to 108). Schroeder (1965) reported that for 106 students classified as psychosomatic, aggressive, exhibiting school difficulties, school phobic, or neurotic, the average IQ was 95.95. Motto and Lathan (1966) studied 47 school-age children in a state hospital and reported that, as a group, they were in the dull normal range of general intelligence. Glavin, Quay, and Werry (1971) reported IQ ranges of 89 to 112 for 11 conduct problem children placed in special classrooms. Fuller and Goh (1981) examined 38 learning disabled and 42 emotionally disturbed public school children and reported lower average IQ scores for the LD than for the ED students (86.13 and 89.50, respectively). As recently as 1983, Forness, Bennett, and Tose reported that 92 subjects (23 girls and 69 boys) who had been inpatients at a neuropsychiatric institute had, on the average, IQ scores in the low 90's.

Reilly, Ross, & Bullock (1979) examined the intellectual performance of 177 adjudicated adolescents and reported a mean IQ score of 90.26, a figure consistent with that of a previous investigation (Bullock & Reilly, 1979). In addition, these researchers reported that subjects scored near average on the Picture Arrangement subtest of the Wechsler Intelligence Scale for Children - Revised (WISC-R) which requires visual sequencing of simple stories, but lowest on those verbal subtests which require knowledge of the "outside world": Information, Similarities, Vocabulary. Finally, a relation between IQ performance and violent behavior was not found in this investigation.

Research on intellectual performance of disturbed children reveals that the majority of mildly and moderately disturbed children fall only slightly below average in IQ. These investigations, taken together, appear to suggest that mild academic deficiencies could be predicted on the basis of observed intellectual functioning. Scruggs and Mastropieri (1984) pointed out that IQ scores in combination with achievement test scores can provide information regarding relative discrepancies between ability and academic performance of the behaviorally disordered population. What IQ scores cannot do is describe behaviorally disordered students' actual levels of academic performance. Kauffman (1985), however, does maintain that IQs of disturbed children are the best predictors of future educational achievement. The following section describes investigations of academic functioning.

Achievement

Reading and Arithmetic

Silberberg and Silberberg (1971) reviewed research on school achievement and delinquency. They cited early studies by Lane and Witty (1934), Bond and Fendrick (1936), Sullivan (1927), and Hill (1935) who found that, in general, delinquents were deficient in reading achievement.

Tamkin (1960), whose subjects included 34 children receiving residential treatment for emotional disorders, reported both the arithmetic and reading grade rating to be within the range commensurate with the mean chronological age of the sample. Arithmetic achievement was significantly lower than reading. Data from the Wide Range Achievement Test (WRAT) showed that 32% demonstrated some degree of educational disability, 41% were educationally advanced, and the remaining 27% were at expected grade level.

Stone and Rowley (1964) tested 116 children referred for psychiatric services using the WRAT. The majority of children fell below the expected level of achievement in reading and arithmetic on the basis of both chronological and mental ages. These children also scored significantly lower in arithmetic than reading. In actual grade placement, a larger proportion were in grades below those expected on the basis of chronological age. Likewise, Reilly, Ross, and Bullock (1979) reported that academic performance was deficient in all areas, with arithmetic scores consistently lower than reading. In addition, Reilly et al. (1979) reported that violent offenders had the lowest reading scores. In a related investigation, Bullock and Reilly (1979) reported lower achievement in all content areas on a similar sample of youthful offenders. Additionally, greatest achievement deficiencies were found for male, minority, and older subjects.

Graubard (1964) compared the performance of 21 children in a psychiatric residential treatment center. Using the Metropolitan Achievement Test and the Stanford Achievement Test, he reported severe reading and arithmetic disability by comparing mental age to expected reading and arithmetic achievement. No evidence supporting a significant difference between reading and arithmetic achievement was found.

Schroeder (1965) compared the WRAT scores of 106 students classified as having emotional problems (psychosomatic, aggressive, school difficulties, school phobia, or neurotic personalities). The mean scores were consistently lower in arithmetic than reading in all five categories. The school difficulties category included the lowest mean achievement level in arithmetic and reading. The highest grade equivalent composite mean was reported in the neurotic-psychotic category. Emotionally disturbed children were deficient at all age levels with respect to school achievement. Schroeder concluded that academic disabilities are concomitant with emotional disturbance and vice versa.

Glavin and Annesley (1966) administered the California Achievement Test to 90 normal boys and 130 behaviorally disturbed boys (who were further divided into conduct problem, withdrawn, and inadequacy-immaturity groups) in public school. Their findings showed 81.5% of the BD group were underachieving in reading and 72.3% underachieving in arithmetic. Academic failure can be expected in a high proportion of delinquent or conduct disordered children according to the review of Silberberg and Silberberg (1971); Glavin and Annesley (1966) found no significant differences in performance between the conduct disordered and the withdrawn group.

Motto and Lathan (1966) found no significant difference in the uniformity of achievement in reading and arithmetic of 47 school-age children from a state hospital. The children were below expectations based upon chronological and mental ages. However, they did find more pronounced retardation in males.

Forness, Bennett, and Tose (1983) found similar results comparing 92 children who had been inpatients at a neuropsychiatric institute. Both boys and girls scored below expected levels on the Peabody Individual Achievement Test, although 12 year old boys were lowest in reading recognition and reading comprehension. In a similar investigation (Forness, Frankel, Caldon, & Carter, 1979), 34 hospitalized patients exhibited deficiencies in all academic areas, particularly math and spelling.

Fuller and Goh (1981) compared 38 learning disabled and 42 emotionally disturbed public school children. The Wide Range Achievement Test scores of LD children were lower than those of BD children on reading, spelling, and math. This was not so, however, on the Minnesota Percepto-Diagnostic Test, although no statistical tests were computed on the results.

Harris and King (1982) compared academic achievement of children classified as having learning problems, behavior problems, learning and behavior problems, or "no problems." They studied scores of 242 public school children administered the Science Research Associates (SRA) Achievement Tests. Those children with learning problems scored lower than the children with no problems. Those with behavior problems did not differ from the no problem category on the SRA subtests of Reading, Math, Science, Use of Sources, but did differ from all groups on Language Arts and Social Studies. The learning and behavior problem group performed lower than all groups on the SRA.

Epstein and Cullinan (1983) also found that for 16 matched pairs (IQ, sex, chronological age, ethnicity) of learning disabled and behaviorally disordered public school students, the BD students scored significantly higher than the LD students on all subjects except the general information subtest of the Peabody Individual Achievement Test (cf., Reilly, Ross, & Bullock, 1979) and the math subtest of the Wide Range Achievement Test. These researchers suggested that differential academic programming may be indicated for LD and BD children.

In contrast, Scruggs and Mastropieri (1984) investigated the Stanford Achievement Test scores of 1480 primary grade special education students (619 learning disabled and 863 behaviorally disordered) in several different content areas. They concluded that the LD and BD children were, in fact, very similar with respect to academic performance, with LD children scoring slightly but consistently higher than BD children. No consistent reading-math discrepancy was noted in either population. Also found was the fact that the variability of BD student performance descriptively exceeded that of LD students; thus, a wider range of academic achievement among BD students may be expected.

In contrast to the above studies, one investigation reported results which suggested that BD students do not exhibit academic deficiencies. Graubard (1971) examined the reading achievement and behavior checklist scores of 108 emotionally disturbed children and concluded, "...all groups' reading commensurate with MA and several groups' reading commensurate with CA" (p. 757). Graubard added, however, that academic retardation in his sample was associated with severity of conduct disorders. Unfortunately, no data were offered to support these conclusions.

Spelling

Few studies in subjects other than reading and arithmetic have been conducted. Glavin and DeGirolamo (1966) found differences between withdrawn and conduct disordered students with respect to types of spelling errors. The withdrawn children made significantly more written spelling errors, while the conduct problem children made significantly more refusals (i.e., refused to complete the task). They concluded that children with emotional problems may show patterns of spelling errors which differ both quantitatively and qualitatively from those of normal children. In addition, as mentioned above, Fuller and Goh (1981) found that learning disabled students scored lower than emotionally disturbed students on tests of spelling achievement.

Psycho-Social Functioning and Academic Performance

The present review of previous investigations can offer little evidence that the reported academic deficiencies of BD children are content specific; that is, research findings tend to support the notion that BD students are deficient in all areas of academic functioning, with some individual investigations reporting more serious deficits in math. Research which has examined academic performance in several different areas within one investigation has supported this conclusion (e.g., Scruggs & Mastropieri, 1984). However, several other researchers have investigated the interaction of academic performance and measures of psycho-social functioning. One major purpose of these investigations, described below, is to identify possible causal explanations for academic deficits.

Glueck and Glueck (1950) reported that delinquents exhibited more dislike for school subjects requiring strict logical reasoning and persistency of effort as well as those dependent upon efficient memory skills. This finding may partially explain some of the previous reports of differentially low performance in math. School achievement of the delinquent students was far below that of nondelinquents.

Graubard (1965) found that 35 delinquents incarcerated at a residential treatment center had similar communication patterns to those of non-adjudicated adolescents. The author maintained, however, that deficits were exhibited in the visual-motor channel (integration level). Delinquents also were reported to exhibit deficits in the Auditory Vocal Automatic modality and in directionality. Findings reported in this investigation, however, may be complicated by reliability and validity limitations of the measures administered (i.e., Illinois Test of Psycholinguistic Ability, Harris Test of Lateral Dominance).

Two investigations examined locus of control and academic achievement with BD students. Hisama (1976) compared 48 special education students with learning and behavior problems to 48 nonhandicapped students on a locus of control measure. It was hypothesized that externality may be a factor for low achievement motivation of behaviorally disordered and learning disabled children. Hisama reported that the Children's Locus of Control Scale showed no difference in scores between normals and LD and BD students. It was concluded that the child with learning and behavior problems may not be more externally oriented than the normal child. In a similar study, Perna, Dunlap, and Dillard (1984) found that for 63 males classified as mildly to moderately emotionally disturbed, those students who felt a high degree of self-responsibility for their successes and failures (internality) showed greater academic gains.

Letteri (1979) provided a "Cognitive Profile" associated with low academic achievement and severe behavior problems as a result of research efforts with 200 subjects (some BD, some not). The cognitive processes associated with low achievement were said to include: Simple (vs. cognitive complexity), leveler (vs. sharpener), intolerant for ambiguous information, global or field dependent (vs. analytical way of perceiving), broad (vs. narrow inconclusiveness in breadth of categorization), non-focuser, and impulsive (vs. reflective).

Four recent studies investigated attitudes and responses to achievement tests themselves. Scruggs, Mastropieri, Tolfa, and Jenkins (1985) examined attitudes expressed by BD students toward the test-taking experience. When surveys were administered at the beginning of the school year, reported attitudes of BD and more average students were very similar. When administered immediately after three days of testing, however, BD students reported more negative attitudes than their regular class counterparts. Taking a different perspective, Forness and Dvorak (1982) examined the general question of academic performance of disturbed or behaviorally disordered students under different testing conditions. Forty adolescents who had been inpatients at a neuropsychiatric institute were tested using the Comprehensive Test of Basic Skills under untimed conditions. Their scores were compared with scores obtained at the end of the normal time limits of the test. The only performance to increase under untimed conditions was that of reading comprehension. Similarly, Scruggs and Mastropieri (in press) trained a sample of mildly handicapped students, mostly BD, on test-taking skills and reported a significant performance advantage on reading subtests. This finding suggests that BD students may be deficient with respect to test-taking skills. In a more recent study, Scruggs, Mastropieri, and Tolfa (1985) reported that test-taking skills training of BD students had differentially raised scores on a "math concepts" subtest over those of LD students to the extent that trained BD students gained 16 percentile points over their untrained counterparts. This finding may help explain why BD students' achievement scores in math are often differentially low.

Conclusions

The investigations reviewed in this paper represent a wide range of populations, all considered in some way "behaviorally disordered." Different assessment measures have been used in a wide variety of different settings. In spite of the diversity of methods, measures, and population samples, however, some broad conclusions can be drawn and are given below.

First, BD students consistently have been seen to exhibit academic and intellectual deficiencies. Although several investigations have examined the possibility of specific content area deficiencies, all evidence to date indicates that academic deficiencies exhibited by this population are global, with a smaller set of investigators suggesting arithmetic performance may be relatively lower than reading. In addition, deficiencies in academic areas have typically been greater than intellectual deficiencies. Investigators who examined ability/performance discrepancies in BD children have indicated that academic achievement is generally below levels predicted by ability tests. These consistent results suggest that the need for academic remediation in this population is as great as the need for behavior management and social skills training.

Whether the reported academic deficiencies of BD students are greater than those typically exhibited by learning disabled students is less certain. Fuller and Goh (1981) and Epstein and Cullinan (1983) reported that LD students scored lower on achievement measures, while Scruggs and Mastropieri (1984) reported that LD students scored consistently higher. In spite of these discrepant findings, however, substantial academic deficiencies have been reported in both populations. In addition, BD students have exhibited consistently higher variability, due no doubt to the fact that LD students are operating under an academic "cut off" level, while BD students are not.

In addition, several variables have been identified which may partially explain observed academic deficiencies. These potentially related variables include attitude toward school subjects (Silberberg & Silberberg, 1971), external locus of control (Hisama, 1976; Perna, Dunlap, & Dillard, 1984), impulsivity (Letteri, 1979), and responses to test-taking situations (Forness & Dvorak, 1982; Scruggs & Mastropieri, in press; Scruggs, Mastropieri, & Tolfa, 1985; Scruggs, Mastropieri, Tolfa, & Jenkins, 1985). Many of these investigations simply describe characteristics of this population, however, and do not provide information that these variables are, in fact, causally related. Further research is needed to document more carefully the reasons for the observed academic deficiencies.

Finally, it must be noted that research concerned with optimal instructional strategies for this population has been greatly neglected, given the nature and extent of the problem. Epstein, Cullinan, and Rose (1980) referred to academic remediation of BD students as an area "... of great concern to special education practitioners, but, ironically, of less concern to researchers" (p. 64). They described the several investigations which had been conducted, virtually all of which examined the role of token reinforcement in increasing academic performance. Although some initial research has been conducted which appears promising in evaluating the effect of such other instructional variables as corrective feedback (e.g., Polsgrove, Reith, Friend, & Cohen, 1979), increased instructional time (e.g., Reith, Polsgrove, Semmel, & Cohen, 1979), self-management (e.g., Cohen, Polsgrove, & Reith, 1979), peer tutoring (Scruggs, Mastropieri, & Richter, in press), and cooperative vs. competitive learning (Scruggs & Mastropieri, 1985), further research is needed to refine these variables and to identify other variables effective in remediating the serious academic deficits of this population.

References

Bond, G., & Fendrick, P. (1936). Delinquency and reading. Pedagogical Seminar and Journal of Genetic Psychology, 48, 236-243.

Bullock, L. M., & Reilly, T. F. (1979). A descriptive profile of the adjudicated adolescent: A status report. In R. B. Rutherford, Jr., & A. G. Prieto (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 2). Tempe, AZ: Arizona State University, Council for Children with Behavior Disorders.

Cohen, R., Polsgrove, L., & Reith, H. J. (1979). An analysis of the effects of goal setting, self-management, and token reinforcement on oral reading performance of children with learning and behavior disorders. In R. B. Rutherford, Jr., A. G. Prieto, & J. E. McGlothlin (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 3). Tempe, AZ: Arizona State University, Council for Children with Behavior Disorders.

Epstein, M. H., & Cullinan, D. (1983). Academic performance of behaviorally disordered and learning disabled pupils. Journal of Special Education, 17, 303-307.

Epstein, M. H., Cullinan, D., & Rose, T. L. (1980). Applied behavior analysis and behaviorally disordered pupils: Selected issues. In L. Mann & D. A. Sabatino (Eds.), The fourth review of special education. New York: Grune & Stratton.

Forness, S. R., Bennett, M. A., & Tose, B. A. (1983). Academic benefits in emotionally disturbed children revisited. Journal of the American Academy of Child Psychiatry, 22, 140-144.

Forness, S. R., & Dvorak, R. (1982). Effects of test time-limits on achievement scores of behaviorally disordered adolescents. Behavioral Disorders, 7, 207-212.

Forness, S. R., Frankel, F., Caldon, P. L., & Carter, M. J. (1979). Achievement gains of children hospitalized for behavior disorders. In R. B. Rutherford, Jr., A. G. Prieto, & J. E. McGlothlin (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 3). Tempe, AZ: Arizona State University, Council for Children with Behavior Disorders.

Fuller, G. B., & Goh, D. S. (1981). Intelligence, achievement, and visual-motor performance among learning disabled and emotionally impaired children. Psychology in the Schools, 18, 261-268.

Glavin, J. P., & Annesley, F. R. (1966). Reading and arithmetic correlates of conduct-problem and withdrawn children. The Journal of Special Education, 5, 213-219.

Glavin, J. P., & DeGirolamo, G. (1966). Spelling errors of withdrawn and conduct problem children. The Journal of Special Education, 4, 199-204.

Glavin, J P., Quay, H. C., & Werry, J. S. (1971). Behavioral and academic gains of conduct problem children in different classroom settings. Exceptional Children, 37, 441-446.

Glueck, S., & Glueck, E. (1950). Unraveling juvenile delinquency. Cambridge, MA: Harvard University Press.

Graubard, P. S. (1971). The relation between academic achievement and behavior dimensions. Exceptional Children, 37, 755-757.

Graubard, P. S. (1965). Psycholinguistic correlates of reading disability in disturbed delinquent children. The Journal of Special Education, 1, 363-368.

Graubard, P. S. (1964). The extent of academic retardation in a residential treatment center. Journal of Educational Research, 58, 78-80.

Harris, W. J., & King, D. R. (1982). Achievement, sociometric status, and personality characteristics of children selected by their teachers as having learning and/or behavior problems. Psychology in the Schools, 19, 452-457.

Hill, G. E. (1935). Educational attainment of young men offenders. The Elementary School Journal, 36, 53-58.

Hisama, T. (1976). Achievement motivation and the locus of control of children with learning disabilities and behavior disorders. Journal of Learning Disabilities, 9, 58-63.

Kauffman, J. M. (1985). Characteristics of children's behavior disorders (3rd ed.). Columbus, OH: Merrill.

Lane, H. A., & Witty, P. A. (1934). The educational attainment of delinquent boys. The Journal of Educational Psychology, 25, 695-702.

Letteri, C. A. (1979). The relationship between cognitive profiles, levels of academic achievement and behavior problems. In R. B. Rutherford, Jr., A. G. Prieto, & J. E. McGlothlin (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 2). Tempe, AZ: Arizona State University, Council for Children with Behavior Disorders.

Messer, S. B. (1976). Reflection-impulsivity: A review. Psychological Bulletin, 83, 1026-1052.

Motto, J. J., & Lathan, L. (1966). An analysis of children's educational achievement and related variables in a state psychiatric hospital. Exceptional Children, 32, 619-623.

Perna, S. J., Dunlap, W. R., & Dillard, J. W. (1984). The relationship of internal locus of control, academic achievement, and IQ in emotionally disturbed boys. Behavior Disorders, 9, 36-42.

Polsgrove, L., Reith, H. J., Friend, M., & Conen, R. (1979). An analysis of the effects of various instructional procedures on the oral reading performance of high school special education students. In R. B. Rutherford, Jr., A. G. Prieto, & J. E. McGlothlin (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 3). Tempe, AZ: Arizona State University, Council for Chilcren with Behavior Disorders.

Reilly, T. F., Ross, D., & Bullock, L. M. (1979). The contemporary adolescent delinquent: Intellectual or impulsive? In R. B. Rutherford, Jr., A. G. Prieto, & J. E. McGlothlin (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 3). Tempe, AZ: Arizona State University, Council for Children with Behavior Disorders.

Reith, H. J., Polsgrove, L., Semmel, M., & Cohen, R. (1979). An experimental analysis of the effects of increased instructional time on the academic achievement of a "behaviorally disordered" high school pupil. In R. B. Rutherford, Jr., A. G. Prieto, & J. E. McGlothlin (Eds.), Monograph in behavioral disorders: Severe behavior disorders of children and youth (Vol. 3). Tempe, AZ: Arizona State University, Council for Children with Behavior Disorders.

Schroeder, L. B. (1965). A study of relationships between five descriptive categories of emotional disturbance and reading and arithmetic achievement. *Exceptional Children*, *32*, 111-112.

Scruggs, T. E., & Mastropieri, M. A. (1985). Competitive vs. cooperative performances of behaviorally disordered American Indian adolescents. *Journal of Instructional Psychology*, *12*, 31-33.

Scruggs, T. E., & Mastropieri, M. A. (in press). Improving the test-taking skills of behaviorally disordered and learning disabled students. *Exceptional Children*.

Scruggs, T. E., & Mastropieri, M. A. (1984). *Academic characteristics of behaviorally disordered and learning disabled students*. Unpublished manuscript, Utah State University.

Scruggs, T. E., Mastropieri, M. A., & Richter, L. L. (in press). Peer tutoring with behaviorally disordered students: Social and academic benefits. *Behavioral Disorders*.

Scruggs, T. E., Mastropieri, M. A., & Tolfa, D. (1985). *The effects of coaching on the standardized test performance of behaviorally disordered and learning disabled children*. Unpublished manuscript, Utah State University, Logan, UT.

Scruggs, T. E., Mastropieri, M. A., Tolfa, D., & Jenkins, V. (1985). Attitudes of behaviorally disordered students toward tests. *Perceptual and Motor Skills*, *60*, 467-470.

Silberberg, N. E., & Silberberg, M. C. (1971). School achievement and delinquency. Review of Educational Research, 41, 17-32.

Stone, F. B., & Rowley, V. N. (1964). Educational disability in emotionally disturbed children. Exceptional Children, 30, 423-426.

Sullivan, E. B. (1927). Age, intelligence, and educational achievement of boys entering Whittier State School. Journal of Delinquency, 11, 23-38.

Tamkin, A. S. (1960). A survey of educational disability in emotionally disturbed children. Journal of Educational Research, 53, 313-315.

Author Notes

Preparation of this manuscript was supported in part by a grant from the Department of Education, Special Education Programs, #G008300008. The authors would like to thank Ursula Pimentel for her assistance in the preparation of this manuscript. Address requests for reprints to Margo A. Mastropieri, Ph.D., Department of Special Education, Utah State University, Logan, UT 84322.

Table 1

BD Academic Characteristics Studies

AUTHORS	SUBJECTS	TASK	RESULTS
Bullock & Reilly (1979)	188 adolescents adjudicated for behavioral offenses.	Wechsler Intelligence Scale, Wide Range Achievement Test (WRAT).	1. Average IQ of 90. 2. Average achievement deficit in all areas. 3. Discrepancies were greatest for males, minorities, older students
Epstein & Cullinan (1983)	16 matched pairs (IQ, sex, CA, ethnicity); LD & BD; public school students	Peabody Individual Achievement Test (PIAT) and Wide Range Achievement Test (WRAT) were administered to both groups.	1. BD students scored significantly higher than LD students on all subjects except general information subtest of PIAT and math subtest of WRAT.
Forness, Bennett, & Tose (1983)	23 girls, and 69 boys who had been inpatients at a neuropsychiatric institute; mean age 10.1 years	Peabody Individual Achievement Test (PIAT) and Wechsler Intelligence Scale for Children-Revised (WISC-R) were administered to all students.	1. Both girls and boys scored below expected levels on PIAT (moderately). 2. Both girls & boys IQ in low 90's 3. 12 yr. old boys worse in reading recognition and reading comprehension. 4. 10 yr. old girls 2.1 yrs. below grade level. 5. 12 yr. old girls 1.7 yrs. below grade level.
Forness & Dvorak (1982	40 BD adolescents (15 males, 25 females) who had been inpatients at a neuropsychiatric institute; mean age 15.7 years	Comprehensive Test of Basic Skills (CTBS) was administered and scored under times and untimed testing conditions.	1. No significant test score differences, except on the reading comprehension subtest.
Forness, Frankel, Caldron & Carter (1979)	34 children (CA 7.0 to 12.9) hospitalized for severe behavior disorders	Peabody Individual Achievement Test (PIAT).	1. Students were deficient in all academic areas, particularly math and spelling. 2. Longer hospitalization periods were associated with greater academic gains.
Fuller & Goh (1981)	38 LD and 42 ED Children; public school setting; mean age 10 years.	Wechsler Intelligence Scale for Children-Revised), Wide Range Achievement Test (WRAT), and Minnesota Percepto-Diagnostic Test (MPD) were administered to all students.	1. Discriminant analysis procedures indicated that LD students & ED students could be accurately placed. 2. LD's lower than ED on IQ, reading, spelling, and math, but not on MPD (however, no statistical tests computed on results).

(table continues)

AUTHORS	SUBJECTS	TASK	RESULTS
Glavin & Annesley (1966)	130 BD boys and 90 normal boys in public school settings. (BD further divided into conduct problem, withdrawn, & inadequacy-immaturity groups).	California Achievement Test (CAT) and Behavioral Scales (Quay & Peterson 67).	1. 81.5% of the BD group were underachieving in reading. 2. 72.3% of the BD group were underachieving in arithmetic. 3. No significant differences in performance were found between the conduct disordered group & the withdrawn group.
Glavin & Degirolamo (1966)	1. 9 ED and 9 Regular Education students; public school setting. 2. 15 ED students classified as either conduct disordered or withdrawn, and reg. ED students.	Spelling words from GATE's *A List of Spelling Difficulties in 3876 words* (1937) were administered to both groups.	1. ED students made more "internal" errors and fewer "external" errors than regular students. 2. Withdrawn students wrote significantly more unrecognizable words. 3. Conduct disordered students made significantly more "refusal" errors.
Glavin, Quay, & Werry (1971)	Conduct problem children placed in experimental special classrooms; 50% Afro-American; IQs 89-112; 1967, N=11, mean age 108 months (age range 91-132); 1968, N=12, mean age 112 months (age range 89-131); both years, N=8.	1967, Wide Range Achievement Test (WRAT); 1968, California Achievement Test (CAT) pre- and post.	1. 1968 arithmetic gain 1.7 years. 2. 1967 arithmetic gain .1 years. 3. 1968 reading gain 1.2 years. 4. 1967 reading gain .5 years. 5. 1968 greater emphasis on academic achievement. 6. Gain indicates program brings changes in specific learning-related behavior and obtains concomitant gains in academic achievement.
Graubard (1971)	108 disturbed students in special schools.	Reading Achievement, Behavior Problem Checklist	1. No overall reading deficiency. 2. Observed deficiencies associated with severity of conduct disorder.
Graubard (1965)	35 disturbed delinquents incarcerated at residential treatment center; age range 8 years 6 months to 10 years 11 months.	Wechsler Intelligence Scale for Children (WISC), Metropolitan Achievement Test (MAT), Illinois Test of Psycholinguistic Abilities (ITPA), Monroe Test of Auditory Blending (MTAB), and Harris Test of Lateral Dominance (HTLD).	1. BD students did not differ from normals in communication pattern. 2. BD students have deficits in the visual-motor channel (the integration level). 3. BD students have deficits in the Auditory Vocal Automatic modality and in directionality.

(table continues)

AUTHORS	SUBJECTS	TASK	RESULTS
Graubard (1964)	21 children in psychiatric residential treatment from 2-8 years (delinquent or neglected); mean age 13 years 10 months (range 10-16); mean grade 7.9 (range 5-11; mean IQ 92.3 (range 71-108); all boys.	Wechsler Intelligence Scale for Children (WISC), Metropolitan Achievement Test, Stanford Achievement Test.	1. Difference between reading and math not significant; mean grade rating both tests 4.75; mean grade reading comprehension 4.87; mean grade arithmetic computation 4.62. 2. Educational disability measured by comparing mental age to reading and arithmetic ages. Severe reading and arithmetic disability found. 3. Not achieving commensurate with mental ages and disabled in academic achievement. 4. No evidence supporting significant difference between reading and arithmetic achievement in population with severe emotional problems over time.
Harris & King (1982)	242 children in grades 4 and 5 in public school settings; students were classified as LP (learning problem N=33), BP (behavior problem N=17), LBP (learning & behavior problem N=19) or NP (no problem N=173)	Science Research Associates Achievement Tests (SRA), Children's Personality Questionnaire (CPQ), L-J Sociometric Test (L-JST).	1. LP students achieved lower scores on SRA, were less preferred by peers, were less intelligent than NP and less assertive than BP and LBP groups. 2. BP did not differ from NP on SRA subtests: Reading, Math, Science, Use of Sources, but did differ from all groups on Language Arts and Social Studies. 3. BP did not differ from any group sociometrically. 4. LBP did perform lower than all groups on SRA, were preferred less by all groups.
Hisama (1976)	48 special ed. children with learning and behavior problems; mean CA 108 months (ranges 96-132); public schools; 3rd or 4th graders. 48 normal 3rd or 3th graders; free from learning and behavior problems randomly selected; mean CA 106 months (ranges 90-136).	Children's Locus of Control Scale (CLCS), Coding Test and Digit Symbol Test from WISC, Wechsler Adult Intelligence Scale (WAIS), NIM game (match game).	1. No significant difference in CLCS scores between normals and LD and BD. BD not externally oriented. 2. Coding Test showed children with internality performed better than those with externality. 3. Within experimental group, externally-oriented child responded to success experience positively and performance depressed under failure condition.
Letteri (1979)	200 subjects (some BD some not).	Cognitive Profile.	1. Cognitive profile associated with low academic achievement & and severe behavior problems is: simple, leveler, intolerant for ambiguous information, global, broad, non-focuser, and impulsive.

(table continues)

AUTHORS	SUBJECTS	TASK	RESULTS
Motto & Lathan (1966)	School-age population of state hospital; 34 boys, mean age 13 years 1 mo. (range 10-2 to 16-9); 13 girls, mean age 11 years 2 mo. (range 9-3 to 15-1); as group, in dull normal of general intelligence.	Wechsler Intelligence Scale for Children (WISC); Wechsler Adult Intelligence Scale (WAIS), Stanford-Binet, Form L; California Achievement Test (CAT), reading and arithmetic.	1. Uniformity of achievement in reading and arithmetic - not significantly different. 2. Females, CA 1.4 below expectations in reading; CA 1.6 below expectations in arithmetic; MA .7 below expectancy in reading, and .9 below in arithmetic. 3. Males, CA 2.6 below reading expectancy; CA 3.7 below expectancy in arithmetic; MA 1.8 below reading, and 1.9 below arithmetic. 4. More pronounced retardation in males. 5. Children in hospital school in excess of 10 months gained in reading and arithmetic achievement to extent expected for their mental ages.
Perna, Dunlap, & Dillard (1984)	63 males classified as mildly to moderately ED in public schools; age range 10-15 years (mean age 12.9 years).	Intellectual Achievement Responsibility (IAR), Chronological age, Stanford-Binet IQ (S-BIQ) or WISC-R, California Achievement Test (CAT)	1. ED students who felt a high degree of self-responsibility for their successes and failures showed greater academic gains.
Reilly, Ross, & Bullock (1979)	177 adolescents adjudicated for specific behavioral offenses.	Wechsler Intelligence Scale for children (WISC-R) Wide Range Achievement Test (WRAT)	1. Average WISC-R IQ of 90.26. Near average scores on Picture Arrangement; lowest scores on Information, Comprehension, Vocabulary. 2. Average achievement was deficient in all areas. Arithmetic scores were consistently lower than reading; violent offenders had the lowest reading scores. 3. A relation between IQ and violent behavior was not found.
Schroeder (1965)	106 students classified in one of five categories (psychosomatic, aggressive, school difficulties, school phobia, neurotic-psychotic personalities); mean age 147.06 months.	Wechsler Intelligence Scale for Children (WISC), Jastak Wide Range Achievement Arithmetic, Jastak Wide Range Achievement Reading (WRAT).	1. Mean scores consistently lower in arithmetic than reading in all five categories. 2. School difficulties category lowest mean achievement level in arithmetic and reading. 3. Highest grade equivalent composite mean in neurotic-psychotic category. 4. Emotionally disturbed children were retarded from age level in school achievement. 5. Educational disabilities concomitant with emotional disturbance and vice versa.

(table continues)

AUTHORS	SUBJECTS	TASK	RESULTS
Scruggs & Mastropieri (in press)	50 BD and 28 LD students in grades 3 - 4.	Training test-taking skills relevant to the Stanford Achievement Test (SAT), reading subtests.	1. BD and LD students exhibited deficiencies on the SAT reading subtests. Test scores improved significantly with training.
Scruggs & Mastropieri (1984)	1480 LD and BD students in grades 1 - 3.	Stanford Achievement Test, all subtests.	1. Only slight differences between Ld and BD groups, with LD students consistently higher in achievement. 2. Factor score patterns of LD and BD students were equivalent.
Scruggs, Mastropieri, & Tolfa (1985)	41 LD and 44 BD students in grades 4-6.	Training test-taking skills relevant to the SAT. reading, and math subtests.	1. Trained LD and BD students gained on the reading decoding subtest relative to controls. 2. Differential gain on the part of trained BD students over trained LD students on "math concepts" subtest.
Scruggs, Mastropieri, Tolfa, & Jenkins (1985)	37 BD students and 50 nonhandicapped students, grades 5 - 6.	Test Attitude Scale (TAS).	1. BD and nonhandicapped students did not differ at the beginning of the school year. 2. After three days of testing, BD students reported lower attitudes in personal feelings and personal importance of tests, but did not differ with respect to attitudes concerning fairness of tests.
Stone & Rowley (1984)	82 boys and 34 girls; mean age 12 years; mean IQ 96.52 (range 62-135).	Wide Range Achievement Test (WRAT), arithmetic and reading parts; Wechsler Intelligence Scale for Children (WISC).	1. In reading and arithmetic, majority of children fell below level of achievement expected on basis of chronological age. 2. In using mental ages as basis for determining achievement level, majority fell below expected level in both reading and arithmetic. 3. Emotionally disturbed children lower in arithmetic scores than reading scores (significantly). 4. In actual grade placement, larger proportion were in grades below that expected on basis of CA.

(table continues)

AUTHORS	SUBJECTS	TASK	RESULTS
Tamkin (1960)	Children receiving residential treatment for emotional disorders in psychiatric hospital; 22 boys, mean age 8.7 years; 12 girls, mean age 9.4 years; combined mean age 9.0 years.	Wide Range Achievement Test (WRAT) arithmetic and reading parts.	1. Both arithmetic and reading grade rating within range commensurate with mean CA of sample. 2. Difference between grade ratings for reading and arithmetic was significant at .005 point based upon one-tailed test (t=2.91). 3. 32% (n=11) demonstrated some degree of educational disability. 41% (n=14) were educationally advanced, and remaining 27% (n=9) were at expected grade level - observing difference between CA and grade rating.

APPENDIX R

Academic Characteristics of Behaviorally Disordered and Learning Disabled Students

Thomas E. Scruggs and Margo A. Mastropieri

Utah State University

Running head: ACADEMIC CHARACTERISTICS

Abstract

The academic performance of 1480 behaviorally disordered (BD) and learning disabled (LD) children attending grades 1-3 was compared. Results indicated that differences in academic performance between BD and LD students was trivial. In addition, supplementary analyses indicated that the two groups did not differ with respect to factor structure of achievement test performance, nor did they differ with respect to reading/math correlations. Implications with respect to cross-categorical education are discussed.

Academic Characteristics of Behaviorally Disordered and Learning Disabled Students

The issue of cross-categorical versus categorical placement in special education has been hotly debated in past years (e.g., Hallahan & Kauffman, 1976; Hewett & Forness, 1974; Heward & Orlansky, 1980). This issue is based in part upon a presumed similarity of academic functioning among children representing different categories of exceptionality. That is, if students of different classifications are to be taught in the same classroom, they should first be shown to be functioning on similar academic levels. However, if students classified as behaviorally disordered (BD) can be shown to be functioning on an academic level different from their learning disabled (LD) counterparts, then cross-categorical placement may be less defensible. If, on the other hand, LD and BD children function on similar academic levels, different arguments against cross-categorical placements must be voiced.

Recently, Epstein and Cullinan (1983) argued convincingly that the level of academic functioning of behaviorally disordered students was, in fact, significantly higher than that of corresponding learning disabled students. These researchers matched 16 pairs of learning disabled and behaviorally disordered students for chronological age, IQ, sex, and ethnicity, and administered to all students several achievement measures. They concluded that with chronological age and IQ so matched, BD students were significantly higher than LD students in all subtests with the exception of the General Information subtest on the Peabody Individual Achievement Test and the Math subtest of the Wide Range Achievement Test (BD students, however, had scored significantly higher on the Mathematics subtest of the PIAT). These significant differences amounted to over a one-year difference in grade level scores, leading authors to suggest that "such differences could present problems related to grouping and other instructional considerations" (Epstein & Cullinan, 1983, p. 305). They concluded, "these data give no support to the supposition that the traditional categories of mild-moderate educational handicaps are highly similar on the characteristics of academic achievement" (p. 305).

The results of the Epstein and Cullinan investigation provide valuable information regarding relative achievement discrepancies of BD and LD students. Some limitations of that study, however, have been noted by the authors. These include, among other things, the facts that relatively small samples of students were employed and that no girls or minority pupils were included in the sample. To these above stated limitations could be added another: the conclusions of Epstein and Cullinan refer to only a small sample of LD and BD students matched on IQ, and provide little information concerning academic achievement levels of large numbers of such students actually enrolled in public school special education classes.

The use of IQ data in investigating the academic characteristics of behaviorally disordered students has been employed frequently in the past (Forness, Bennett, & Tose, 1983; Graubard, 1964, 1971; Motto & Wilkins, 1968). Kauffman (1981) has indicated that use of IQ data on behaviorally disordered students is critical for effectively assessing the academic characteristics of this population. Although matching on IQ with behaviorally disordered and other populations does provide information regarding relative discrepancies between ability and academic performance of the behaviorally disordered population, it does not describe the actual level of academic performance exhibited by behaviorally disordered students actually enrolled in special education classes and how this performance differs from that of their learning disabled counterparts. The Epstein and Cullinan (1983) study is most informative regarding the relative ability/academic performance discrepancy of their sample of the two populations, but provides little information regarding the direct comparison of learning disabled and behaviorally disordered students on measures of academic functioning. The present investigation was intended to investigate this issue by examining the achievement test scores of a large sample of LD and BD children as they were enrolled in special education classrooms. Through this procedure, it was thought that evidence could be acquired regarding possible academic differences in performance between these two populations.

Method

Data were collected from 1480 students in grades 1-3 attending special education classrooms in 58 elementary schools in a western metropolitan area. Of this population, 95% were Anglo, and 5% represented minority groups including Black, Hispanic, and Native American; 68.3 percent (1012) were males, and 31.3% (470) were females. Three hundred eighty-two students were attending first grade, 529 students were attending second grade, and 571 students were attending third grade. Six hundred nineteen (42%) were classified as LD and 863 (58%) were classified BD according to Public Law 94-142 and local criteria. These criteria included, for LD students, average or above intelligence and a 40% discrepancy between ability and achievement in two areas of academic functioning. Criteria for classification as behaviorally disordered included average or above intelligence and marked deficits in behavioral and/or emotional functioning documented by teacher and psychologist, and which had proven resistant to simpler remediation. No academic criteria were specified for BD students. One thousand, three hundred and forty-seven students (91%) were attending resource room placements, while 135 students

(9%) were attending self-contained classrooms. IQ data for this population were not available and, in fact, were not solicited for the purposes of this study. Data were collected on the subjects for subtests of the 1973 edition of the Stanford Achievement Test (SAT) (Madden, Gardner, Rudman, Karlsen, & Merwin, 1973). All test data were collected from the same administration, spring, 1983.

Results

Main Analyses

Multivariate analysis of variance (MANOVA) tests were computed between groups at each grade level, with raw scores from the SAT subtests as dependent measures. The MANOVA procedure was used to take into account the high level of intercorrelations between subtests, and to control for an inflated experiment-wise alpha level thought likely to result from repeated t tests on non-independent comparisons (Bock & Haggard, 1968; Kerlinger & Pedhazur, 1973; Levin, in press; Marascuilo & Levin, 1983; Winer, 1971). Raw scores, rather than grade equivalents or percentiles, were computed because the ratio nature of the numbers was more appropriate for meeting the assumptions of analysis of variance (Ferguson, 1968), and because raw scores provide a more precise measure of test behavior.

Analysis of the data revealed a significant multivariate "F" approximation of 5.34, $p < .001$ for second graders, a significant multivariate "F" approximation of 2.20, $p < .033$ for third graders, and a nonsignificant multivariate "F" approximation of .87, $p < .48$ for first graders. Visual inspection of the descriptive data presented in Table 1 indicates that the achievement scores consistently favor the LD group over the BD group, although the effect sizes are small enough in all cases to constitute questionable practical educational importance (Total Score effect sizes of .14, .18, and .08 for first, second, and third graders, respectively). As seen in Table 2, these differences rarely exceed three or four months in grade equivalent scores.

Insert Tables 1 and 2 about here

The finding of a nonsignificant multivariate effect in the first grade sample precluded further analysis with univariate tests (Marascuilo & Levin, 1983). However, univariate t tests were computed on the second and third grade levels, for which significant multivariate effects had been found. To control for the possibility of Type I errors, specific pairwise comparisons were made at a level of significance appropriate to a familywise alpha level of .05 for each grade level.[1] In the case of the seven subtests on the second and third grade level, the resulting alpha was .007. By these rather rigid criteria, significant differences favoring the LD group were nonetheless found at the second grade level for the Vocabulary, Listening Comprehension, Social Science, and Science subtests. Differences betwen groups in Total Math and Spelling approached significance, but not at the level required by this analysis. Differences in reading were negligible, $\underline{t} < 1$ in absolute value. At the third grade level, no comparisons approached significance at the required level, and four of the seven comparisons resulted in $\underline{t}$'s < 1 in absolute value. The fact that a significant multivariate effect but no univariate effects were found is not uncommon and is doubtless a result of the fact that the MANOVA takes into account the high level of correlations between subtests, while the univariate tests do not (Winer, 1971).

Supplementary Analysis

Since statistical differences between BD and LD students were seen to be few, resulting in small effect sizes, supplementary analyses were computed to determine whether the patterns of achievement test performances could be seen to be different for the two groups. To this end, separate factor analyses were computed for BD and LD students at each grade level in order to determine whether the groups differed from each other with respect to underlying factor structure. Each of the six separate factor analyses revealed only one factor, accounting for between 81 to

88% of total variance, and indicating that over all subtests, only one factor was being measured for each group (perhaps, a "general cognitive ability" factor), and that no difference in factor structure between BD and LD groups was discernible. In a follow-up analysis, individual correlations were computed between Total Reading and Total Math subtests for BD and LD students at each grade level. Resulting correlations ranged from .78 to .88 (all p's $< .01$) for all groups. Comparisons made via Fisher's Z transformations (Ferguson, 1981) at each grade level indicated that at no point were correlations for BD students statistically different from correlations for LD students (all p's $> .20$).

Discussion

Results of the present investigation suggest that BD students do not show better academic performance than do their LD age peers when academic achievement scores of students actually attending special education placements are examined. These findings are in sharp contrast with those of Epstein and Cullinan (1983) who suggested that academic performance of BD students is typically higher than that of LD age peers. The reason for these discrepant findings very likely has to do with the fact that the Epstein and Cullinan subjects were matched by IQ, while the subjects in the present investigation represented the total number of a sample of students enrolled in LD and BD classes without respect to intellectual functioning. While the findings of

Epstein and Cullinan are of theoretical importance in that they underline differences in performance discrepancies between the two populations in the sample selected, they do not provide direct evidence concerning how a large sample of these students actually functions in classes compared with their learning disabled counterparts. The conclusions of the present research indicate that at least at the primary grade levels in the population sampled, LD and BD children are in fact very similar with respect to academic performance. Even though statistically significant differences were found on some comparisons, it must be remembered that the large sample size resulted in sufficient statistical power to discern relatively small effect sizes (Cohen, 1968). In fact, for Total Reading, Total Math, or Total Battery scores, these differences do not exceed two months in grade equivalent scores.[2]

Although the sample size used in this investigation was relatively large, it should be recalled that the subjects came from only one geographical area. This fact may present problems in generalization of findings. However, it must also be maintained that the standards for inclusion in special education placement in this area are very similar to criteria used around the country. In fact, these criteria make the findings more surprising in that specific ability/performance discrepancies in areas of academic functioning are necessary requirements for LD placement, while they are not for BD placement. Nevertheless, the strong similarities between the two groups indicate that, for one reason or another, many LD and BD students in the primary grades apparently do function on a highly similar academic level. This finding does not support the assertion of Cullinan, Lloyd, and Epstein (1981) that academic deficits may be minimal in the primary grades and increase with age. It was found, however, that the variability of BD student performance descriptively exceeded that of LD students at all grade levels. Such higher levels of variability on the part of BD students have been reported by Forness et al. (1983). Although the relatively higher descriptive level of variability here may simply be an artifact of the fact that an academic cutoff level was operating for LD but not BD students, it does suggest that a special education teacher may expect to find a wider range of academic achievement among BD students.

In contrast to the Epstein and Cullinan (1983) investigation, no evidence is given by these data that academic programming should proceed differentially for the two groups. However, the fact that two groups are functioning at a similar academic level does not necessarily mean that instructional procedures should be the same. It may be, for example, that the BD group may be more responsive to token economies and direct instruction in independent study strategies, while the LD group may be more responsive to peer tutoring and small-group teacher-led direct instruction procedures. At present, however, it must be concluded that little is known about optimal instructional strategies for LD vs. BD children, and it is the opinion of the present authors that research is greatly needed in this area.

The reason these two supposedly discrepant groups function in such a similar level of academic performance is uncertain, and cannot be given on the basis of the data presented here. It has often been stated in practice by those who work with LD and BD children that the causal link between behavior problems and learning disabilities is a strong one whose directionality is often in question. It may be that the causal relation between learning and behavioral disabilities is of sufficient strength that academic shortcomings are a frequent consequence, regardless of the nature of special education classification.

In spite of the apparent discrepancies between the present investigation and the Epstein and Cullinan (1983) study, the authors would like to end on a note of concordance with those researchers. In our view, Epstein and Cullinan are quite correct in their assertion that effectiveness of service is a much higher priority than the categorical versus cross-categorical nature of that service, an assertion for which empirical support is available (Heller, Holtzman, & Messick, 1982). Although the present data suggest that cross-categorical placement may be advisable, the present authors would rather see effective educational programming in categorical settings than ineffective teaching in cross-categorical settings. It is thought, however, that the search for optimal educational settings can parallel the search for optimal educational strategies within such settings, and it is to these ends that the present research was addressed.

References

Bernhardson, C. S. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. Biometrics, 31, 229-232.

Boch, R. D., & Haggard, E. A. (1968). The use of multivariate analysis in behavioral research. In D. K. Whitla (Ed.), Handbook of measurement and assessment in behavioral sciences. Reading, Mass.: Addison-Wesley.

Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. Journal of Special Education, 14, 295-309.

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 68, 66-74.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cullinan, D., Lloyd, J., & Epstein, M. (1981). Behavior disorders of children. Englewood Cliffs, NJ: Prentice-Hall.

Epstein, M. H., & Cullinan, D. (1983). Academic performance of behaviorally disordered and learning disabled pupils. Journal of Special Education, 17, 303-307.

Ferguson, G. A. (1981). Statistical analysis in psychology and education (5th ed.). New York: McGraw-Hill.

Forness, S. R., Bennett, M. A., & Tose, B. A. (1983). Academic deficits in emotionally disturbed children revisited. Journal of the American Academy of Child Psychiatry, 22, 140-144.

Graubard, P. S. (1964). The extent of academic retardation in a residential treatment center. Journal of Educational Research, 58, 78-80.

Graubard, P. S. (1971). The relationship between academic achievement and behavior dimensions. Exceptional Children, 37, 755-757.

Hallahan, D. P., & Kauffman, J. M. (1976). Introduction to learning disabilities. Englewood Cliffs, NJ: Prentice-Hall.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.) (1982). Placing children in special education: A strategy for equity. Washington, D.C.: National Academy Press.

Heward, W. L., & Orlansky, M. D. (1983). Exceptional children. Columbus, Ohio: Merrill.

Hewett, F. M., & Forness, S. R. (1974). Education of exceptional learners. Boston: Allyn & Bacon.

Kauffman, J. M. (1981). Characteristics of children's behavior disorders (2nd ed.). Columbus, Ohio: Merrill.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart, & Winston.

Levin, J. R. (in press). Some methodological and statistical "bugs" in research in children's learning. In M. Pressley & C. J. Brainerd (Eds.), The cognitive side of memory development. New York: Springer-Verlag.

Madden, R., Gardner, E. F , Rudman, H. C., Karlsen, B., & Merwin, J. C. (1973). Stanford Achievement Test. New York: Harcourt Brace Jovanovich.

Marascuilo, L. A., & Levin, J. R. (1983). Multivariate statistics in the social sciences: A researcher's guide. Monterey, California: Brooks/Cole.

Motto, J. J., & Wilkins, G. S. (1968). Educational achievement of institutionalized emotionally disturbed children. Journal of Educational Research, 61, 218-221.

White, K. R. (1984). Unpublished data. Utah State University, Logan, Utah.

Winer, B. J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

Footnote

This research was supported in part by a grant from the U.S. Department of Education, Research in the Education of the Handicapped, #G008300008. The authors would like to thank Marilyn Tinnakul and Ursula Pimentel for their assistance in the preparation of this manuscript. Address requests for reprints to Thomas E. Scruggs, Ph.D., Developmental Center for the Handicapped, Utah State University, Logan, Utah 84322.

[1] It can be argued that multiple _t_ tests on non-independent data sets do not inflate the Type I error probability as much in actual practice as expected by statistical theory, and in fact, some recent Monte Carlo studies have supported this argument (Bernhardson, 1975; Carmer & Swanson, 1973; White, 1984). The decision made here was to use the more conservative procedure, especially considering the fact that the large sample size allowed sufficient power to detect relatively small differences even when the pairwise alpha level was quite small.

[2] A case may be made that although academic functioning appears similar given a static achievement test measure, the population may differ with respect to _rate_ of learning. If this were true, however, one would expect the BD students to begin to surpass the LD students academically by the second or third grade. Such differences over grade levels, however, were not observed.

Table 1

Descriptive Data and Statistical Comparisons

	BD (N = 253)		LD (N = 129)			
	Percentile	Grade Equivalent	Percentile	Grade Equivalent	*t**	Effect size
First grade						
Total reading	20	1.4	23	1.5	-$^{+}$	-.09
Total math	18	1.3	24	1.4	-	-.15
Vocabulary	23	1.0	30	1.4	-	-.16
Listening comprehension	16	K-6	22	K-9	-	-.16
Total	12	1.1	18	1.3	-	-.14

	BD (N = 323)		LD (N = 206)			
	Percentile	Grade Equivalent	Percentile	Grade Equivalent	*t**	Effect size
Second grade						
Total reading	26	2.0	28	2.0	.55	-.05
Total math	26	1.9	34	2.1	2.16	-.20
Vocabulary	16	1.8	26	2.2	2.95**	-.27
Listening comprehension	11	1.3	20	1.9	3.31**	-.30
Spelling	12	1.6	16	1.8	1.99	-.18
Social science	14	2.0	28	2.2	3.43**	-.31
Science	14	1.5	24	2.1	3.10**	-.29
Total	16	1.6	26	1.8	1.99	-.18

(table continues)

	BD (N = 287)		LD (N = 284)			
	Percentile	Grade Equivalent	Percentile	Grade Equivalent	t*	Effect size
Third grade						
Total reading	23	2.5	24	2.5	.14	-.01
Total math	16	2.9	18	3.0	1.51	-.13
Vocabulary	24	2.5	31	2.9	2.00	-.17
Listening comprehension	20	2.5	24	2.8	1.61	-.14
Spelling	13	2.5	12	2.5	-.57	.05
Social science	22	2.8	22	2.8	.50	.04
Science	12	2.4	16	2.6	-.08	.01
Total	24	2.7	24	2.7	.95	-.08

*All t statistics were computed on raw scores.

**Statistically significant at the pre-specified probability level, $p < .007$.

+Because of a non-significant multivariate effect, univariate statistics were not computed.

APPENDIX S

Improving the Test-Taking Skills of Learning Disabled Students

Thomas E. Scruggs, Ph.D.
Vesna Jenkins
Utah State University

How are your test-taking skills?

1. The short story, "The Four Seasons," is about:
   a. vegetation in North America
   b. wind current and their effects
   c. the changing weather
   d. the growth process

2. The greatest advantage of using slent in the manufacture of steel is that slent makes steel
   a. transparent
   b. stainless
   c. heavy
   d. bulky

3. The Japanese game of paduki
   a. can only be played by the Imperial Family
   b. is sometimes played indoors
   c. can never be played for more than 30 minutes
   d. is always played at every celebration

4. When Bestor crystals are added to water
   a. heat is given off
   b. the temperature of the solution rises
   c. the solution turns blue
   d. the container becomes warmer

The reasoning strategies are explained, followed by the correct answer:

1. The convergence strategy (stem), recently described by Smith (1982), involves teaching test-takers to examine all choices presented after the stem of a multiple-choice question in order to analyze the relationships of the distractors to each other and, thereby, identify the choice most likely to be correct. (1. c).

2. Absurd options can be eliminated as incorrect choices, and thus, increase the probability of choosing the correct answer. (Gibb, 1964). (2. b).

3. Specific determiners (e.g., always, never, all), are words which provide cues to the likely correctness of choices, especially on true/false items. (Slakter, 1970). (3. b).

4. Identifying similar (but slightly different) options again narrows down the possibility of choosing incorrect answers. (Millman, 1969), (4. c).

Should guessing and answer changing be encouraged?

Usually students are advised not to guess on standarized multiple choice tests. However, according to Hammerton (1965) and Bauer (1973), testwise students tend to guess more often than their naive counterparts, and as a result, obtain higher scores. Thus, an appropriate guessing strategy should be employed.

Ebel (1965) concludes from his study with true/false tests that "students seeking highest scores on a test are well advised to answer all questions even when the usual correction is applied (their blind guesses to true/false tend to be correct more than half of the time)."

The problem to solve now becomes "How does a test-taker decide which answer is the best guess?" Numerous testwiseness suggestions are provided by Millman's (1969) and Smith's (1982) guidelines.

Beck (1978) studied the effect of changing item responses on scores of elementary school children on a standardized achievement test. Results clearly indicated that response changes on multiple-choice items tend to improve test scores.

In spite of conventional wisdom regarding guessing and answer changing, research evidence indicates that:

— Students should answer all questions, even when guessing is penalized.

— Students should be encouraged to change any answer they have had second thoughts about.

Do separate answer sheets inhibit the performance of learning disabled students?

Yes, according to a recent study performed at Utah State University, LD and nondisabled students were given three subtests of the Comprehensive Tests of Basic Skills (CTBS) for which correct answers were identified in the test book. Students were instructed to record the correct answers on the separate answer sheet as quickly and efficiently as possible. Learning disabled students' performance was found to be slower, less accurate, and less neat than their nonhandicapped peers. Figure A shows differences between LD and regular classroom students with respect to accuracy and fluency on completion of the separate answer sheet. This discrepancy could contribute to measurement error in the LD population. However, it would also seem that LD students improved appreciably in use of separate answer sheets with practice. Figure B shows increase in fluency and accuracy of LD students after only three practice sessions with teacher feedback.

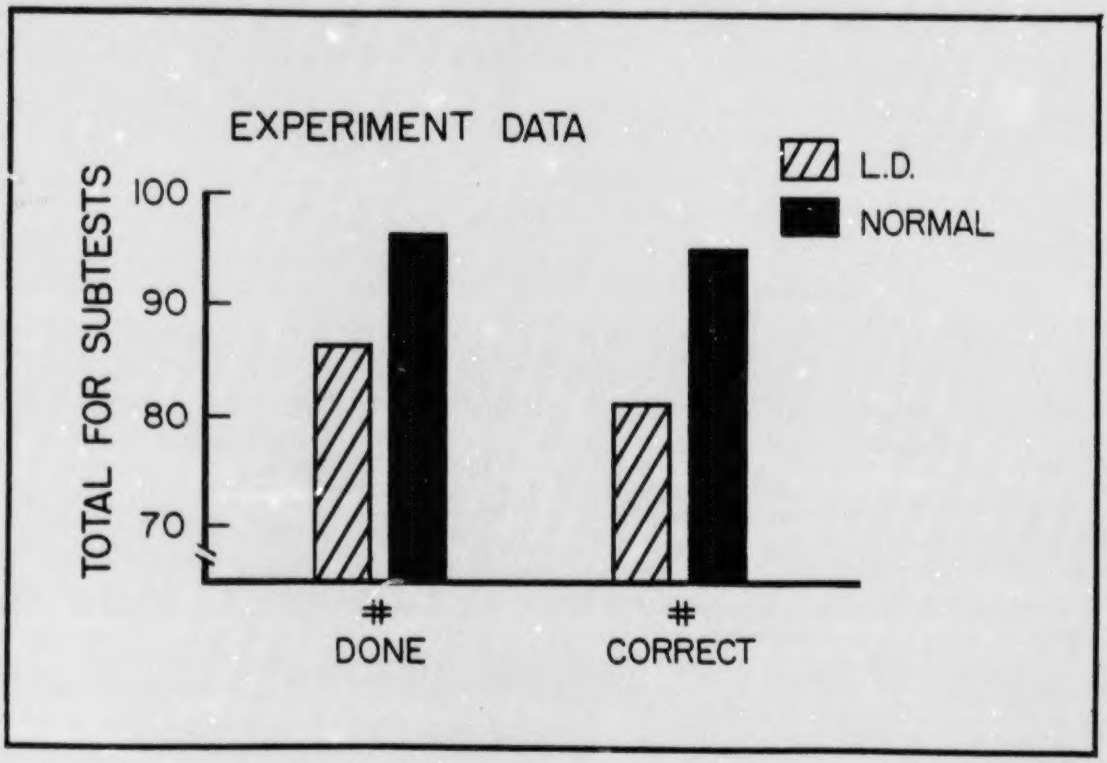


Figure A

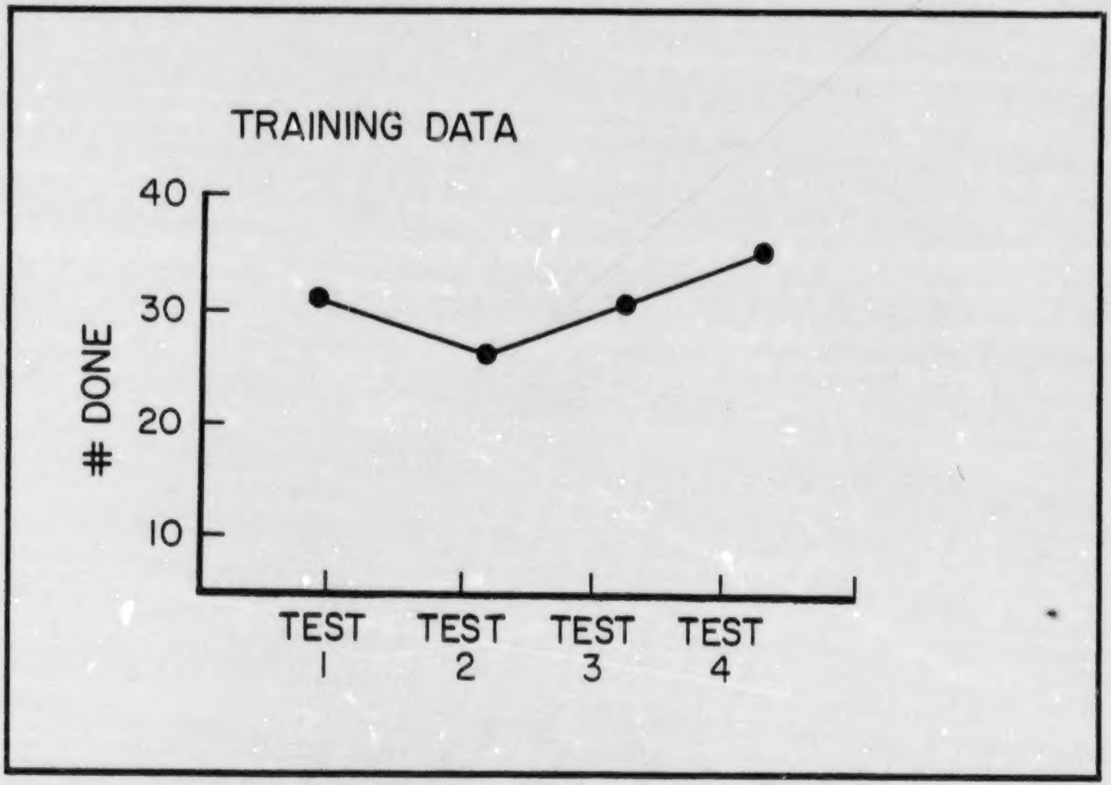


Figure B

Are learning disabled students deficient in test-taking skills? If so, do learning disabled students benefit from training?

Yes, learning disabled students are deficient in test-taking skills. Scruggs (1984, 1985) found LD students differed from their nonhandicapped peers with respect to use of appropriate strategies on standardized achievement tests. These strategy deficits included use of prior knowledge, use of deductive reasoning skills, attention to appropriate distractors, and selection of strategies appropriate to correctly answering different types of items.

Recently, LD students have been trained in using appropriate test-taking strategies. Results indicated that test scores of trained students improved as much as 8-10 percentile points on reading achievement tests over untrained control students (Scruggs & Mastropieri, in press). In addition, a separate investigation revealed that students' attitude toward tests qualitatively improved as a result of training.

What should LD students be taught about test taking?

Our recent research indicates that LD students benefit most from extended, guided practice and general familiarity with test conventions and formats. To this end, LD students should be given relevant practice with questions and formats similar to those which they will see on achievement tests. (Students, of course, should not be given the exact items they will be tested on.)

In addition, the following strategies have been successfully taught to LD students and have been effective in improving test scores:

1. Never skip an answer.

2. Be certain to attend to all distractors and refer to the reading passage, even if you are "very sure" your answer is correct.

3. If you are having great difficulty reading a passage, read the questions and try to answer them anyway. If you have difficulty with some words in the questions, or distractors, answer anyway and base your answers on the words you can read.

4. If you have attended to all parts of a passage and test question and still do not know an answer, there is still a good chance of getting the correct answer if you guess.

5. Be certain you are attending to the appropriate stimulus, such as the underlined sound in a "word study skills" subtest. As in other subtests, wrong answer choices are given which may look correct at first glance.

6. Make sure you answer every item, even if you must hurry and guess a lot near the end. You will probably get some of the answers correct.

Examples and practice activities will help develop these test-taking skills.